

AD-A058 528

MISSOURI UNIV-COLUMBIA TAILORED TESTING RESEARCH LAB

F/G 14/2

A LIVE TAILORED TESTING COMPARISON STUDY OF THE ONE- AND THREE---ETC(U)

JUN 78 W R KOCH, M D RECKASE

N00014-77-C-0097

UNCLASSIFIED

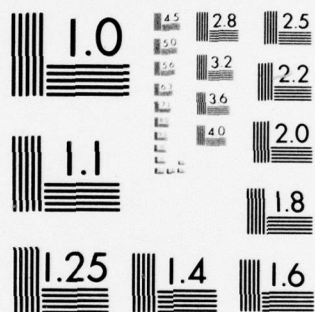
RR-78-1

NL

| OF |  
AD  
A058528



END  
DATE  
FILMED  
11-78  
DDC



MICROCOPY RESOLUTION TEST CHART  
NATIONAL BUREAU OF STANDARDS-1963-A

ADA 058528

AD No. \_\_\_\_\_  
DDC FILE COPY

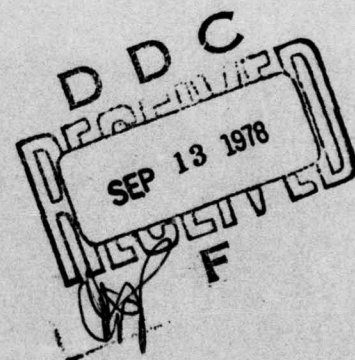
LEVEL II

2  
12

# A Live Tailored Testing Comparison Study of the One- and Three- Parameter Logistic Models

William R. Koch  
and  
Mark D. Reckase

Research Report 78-1  
June 1978



Tailored Testing Research Laboratory  
Educational Psychology Department  
University of Missouri  
Columbia MO 65211



Prepared under contract No. N00014-77-C-0097, NR150-395  
with the Personnel and Training Research Programs  
Psychological Sciences Division  
Office of Naval Research

Approved for public release; distribution unlimited.  
Reproduction in whole or in part is permitted for  
any purpose of the United States Government

78 09 12 009

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

14

RR-78-1

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Research Report 78-1	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Live Tailored Testing Comparison Study of the One- and Three-Parameter Logistic Models.	5. TYPE OF REPORT & PERIOD COVERED Technical Report	
6. AUTHOR(s) William R. Koch & Mark D. Reckase		7. PERFORMING ORG. REPORT NUMBER N00014-77-C-0097
8. CONTRACT OR GRANT NUMBER(s)		9. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS P.E.: 61153N Proj.: RR042- T.A.: 042-04-01 04 W.V.: NR150-395
10. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Educational Psychology University of Missouri - Columbia Columbia, Missouri 65201		11. REPORT DATE June 78
12. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research Arlington, Virginia 22217		13. NUMBER OF PAGES 36
14. MONITORING AGENCY NAME & ADDRESS (If different from Controlling Office) RR04204		15. SECURITY CLASS. (of this report) Unclassified
16. DISTRIBUTION STATEMENT (of this Report) Approval for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.		17. DECLASSIFICATION/DOWNGRADING SCHEDULE
18. SUPPLEMENTARY NOTES DDC SEP 13 1978		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Testing Rasch Model Ability Testing Tailored Testing Latent Trait Models Computerized Testing		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) Although both the one-parameter and the three-parameter logistic models have frequently been applied to computerized tailored tests, no direct empirical comparisons of the two models have been reported in the literature. The major research problem, therefore, was to compare the models based on a live tailored testing study conducted with two identical item pools derived from a vocabulary test. A total of 128 undergraduate		

DD FORM 1 JAN 73 1473

EDITION OF 1 NOV 65 IS OBSOLETE  
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

78 09 12 009 410 502 Lu

cont.

and graduate students enrolled in educational psychology and measurement courses at the University of Missouri - Columbia served as examinees for the study. A counterbalanced test-retest design was employed in which there were two separate test sessions one week apart for each examinee, with both the one- and three-parameter tests administered at each session. The tailored tests were administered on Applied Digital Data Systems Consul 980 cathode ray tube terminals which were connected to an IBM 370/168 computer through a timesharing system. Relative efficiency curves, test-retest reliability coefficients, goodness of fit of the models, convergence rates, descriptive statistics, and the correlation of ability estimates with an outside criterion were used to compare the two models. Test items were selected for administration based on the information function, and maximum likelihood ability estimation was employed. In addition, an attitude survey was constructed to measure several dimensions of student attitudes toward tailored tests. The results of the study indicated an overall superiority for the three-parameter procedure over the one-parameter tailored tests, due to higher reliability, better fit of the model, and greater test information. However, implicit in these results was the assumption that an encountered ability estimate nonconvergence problem could be solved for the three-parameter model. The nonconvergence of ability estimation of the three-parameter model in conjunction with maximum likelihood estimation resulted when the item pool was too difficult for the examinee. The attitude scale results reflected generally favorable student attitudes toward tailored testing.

# CONTENTS

Introduction . . . . .	1
Latent Trait Models . . . . .	2
Method . . . . .	3
Item Calibrations . . . . .	3
Tailored Testing Procedure . . . . .	4
Design . . . . .	6
Sample . . . . .	6
Attitude Survey . . . . .	7
Analyses . . . . .	7
Results . . . . .	9
Goodness of Fit . . . . .	9
Reliability . . . . .	9
Other Correlation Analyses . . . . .	13
Information Function Analyses . . . . .	13
Convergence Plots . . . . .	15
Attitude Scale Characteristics . . . . .	15
Attitude Scale Results . . . . .	19
Summary and Discussion . . . . .	26
Conclusion . . . . .	30
References . . . . .	31
Appendix A: Item Parameter Distributions . . . . .	33
Appendix B: Attitude Survey . . . . .	34
Appendix C: Ability Estimate Distributions . . . . .	36

ACCESSION for	
NTIS	White Section <input checked="" type="checkbox"/>
DDC	Buff Section <input type="checkbox"/>
UNANNOUNCED	<input type="checkbox"/>
JUSTIFICATION	
BY	
DISTRIBUTION/AVAILABILITY CODES	
Dist.	SPECIAL
A	

## A LIVE TAILORED TESTING COMPARISON STUDY OF THE ONE AND THREE PARAMETER LOGISTIC MODELS

Tailored testing derives its name from its primary aim and characteristic, which is to attempt to "tailor" a test for a given individual, often using computer capabilities. That is, rather than administering the same set of test items to all examinees, the tailored testing procedure presents a unique set of items to each examinee which is selected to match item difficulty parameters to a person's estimated ability level. This technique has been developed since "an examinee is measured most effectively when the test items are neither too difficult nor too easy for him" (Lord, 1970, p. 139). Thus, one goal of the tailored testing procedure is to select items from a precalibrated item pool stored in the computer so that the items administered provide a maximum amount of information about a person's ability. In general, tailored testing procedures require three components: a pool of calibrated items, an item selection technique, and a scoring method (Patience, 1977). The main purpose of the present study was to perform an empirical comparison of two different tailored testing models in order to collect evidence for the recommendation of one model over the other in this specific situation.

Several tailored testing procedures have been developed to date (Reckase, 1977). Many of the procedures employ either a one-parameter or a three-parameter logistic model for item calibration and either maximum likelihood or Bayesian ability estimation procedures. Since both the one- and the three-parameter logistic models assume that measurement takes place in a unidimensional latent space, a major long range goal of the present research program is to determine the effects of the violation of this assumption in a tailored testing setting. This is a quite logical endeavor in light of the fact that achievement tests, to which tailored tests are now frequently beginning to be applied, routinely measure several dimensions.

Numerous applications of the one- and three-parameter logistic models have been previously reported in the literature both for group tests and tailored tests (Woodcock, 1973; Jensema, 1974; Ireland, 1976; English, Reckase, & Patience, 1977; Lord, 1977; Marco, 1977; Rentz & Bashaw, 1977). In addition, several model comparison studies have been done, but these usually have been restricted to calibration research (Urry, 1970; Hambleton & Traub, 1971; Reckase, 1977) or have used simulated data (Hambleton, 1969; Reckase, 1972). However, no research has been reported in the literature that directly compares the one- and three-parameter logistic models in a live tailored testing situation with real data. The present study was intended to deal with this issue.

The bases for the comparisons of the two tailored testing procedures will be (a) the goodness of fit of the models using mean squared deviations of observed from predicted response data, (b) the reliabilities of the two methods, (c) the ability estimates yielded by the two procedures, (d) the correlation of the ability estimates with the same outside criteria, (e) descriptive statistics for each procedure, (f) the rates at which the two methods converge to ability estimates, and (g) the information functions for the two procedures.

Before proceeding to an empirical study comparing the two models with tailored achievement tests, it was deemed essential to begin with the more basic level of comparing the two models on live tailored tests using a pool of unidimensional items. At the time of this writing, even this first step in empirical model comparison had not been reported in the literature. Thus the purpose of the present study was to conduct a live tailored testing investigation of the one- and three-parameter logistic models for the unidimensional case as input and groundwork for examining the multidimensional case at a future date. We begin with a brief discussion of the two latent trait models.

### Latent Trait Models

The Rasch (1960), or one-parameter logistic model, is thoroughly described in a recent article by Wright (1977). Here let it suffice to say that the one parameter model requires only one ability parameter,  $\theta_j$ , for each person and one item difficulty parameter,  $b_i$ , for each item to describe the interaction between an examinee and a test item. The exponential form of the simple logistic model is

$$P\{u_{ij}\} = \frac{e^{u_{ij}(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}} \quad (1)$$

where  $u_{ij}$  is the score (0 or 1) on Item  $i$  by Person  $j$ ,  $\theta_j$  and  $b_i$  are as defined above, and  $P\{u_{ij}\}$  is the probability that  $u_{ij}$  equals 0 or 1.

In contrast, the three-parameter logistic model presented by Birnbaum (1968) requires the estimation of three item parameters to describe the interaction between test items and examinees. The model is given by

$$P_{ij} = P\{u_{ij} = 1\} = c_i + (1 - c_i) \frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}} \quad (2)$$

where  $P\{u_{ij} = 1\}$  is the probability of a correct response by Person  $j$  to Item  $i$ ;  $c_i$  is the guessing parameter for Item  $i$ ;  $D$  is a scaling constant equal to 1.7;  $a_i$  is the item discrimination parameter;  $b_i$  is the item difficulty parameter; and  $\theta_j$  is the ability parameter for Person  $j$ .  $Q_{ij}$ , the probability of an incorrect response, is defined simply as  $1 - P_{ij}$ .

Both models have in common the assumption that the items are scored dichotomously, that the latent trait being measured by the items is unidimensional, that the model describes the interaction between a person and an item, and that local independence holds (Lord and Novick, 1968). The graph of the probability of a correct response against ability is called an item characteristic curve (ICC).

# METHOD

## Item Calibrations

The source of items used for the tailored testing comparison study was the Syracuse Adult Development Study vocabulary tests, (Monge & Gardner, 1972), Forms C2, D2, and E. Three successive revisions of these tests were performed during their development based on various item analyses and reliability studies. All of the items were of the multiple choice form with five alternatives per item. Each item had the same stem, "which word below most nearly corresponds in meaning to the word . . .", followed by the word itself and the five alternatives. The result was a set of vocabulary test forms of good quality (KR-20 = .90) for use with adults. Response data were collected from a large, cross-section of adults. A principal components factor analysis of the interitem tetrachoric correlation coefficients conducted on form D-2 indicated that only one factor was present in the test with an eigenvalue greater than 1.0, which accounted for approximately 41% of the variance, with a sample size of 1,000 (Reckase, 1972).

Two identical pools consisting of 72 multiple choice vocabulary items were constructed, one for use with the one-parameter model and the other for the three-parameter model. The one-parameter model pool was calibrated using a modified version of a program given in an article by Wright and Panchapakesan (1969). For the three-parameter pool, the LOGIST program developed by Wood, Wingersky, and Lord (1976) was used. The decision to use these two methods of item calibration was based on the results of a comprehensive review of available calibration procedures reported in a previous technical report (Reckase, 1977).

Table 1 presents the means, standard deviations, and ranges of the item parameter estimates resulting from the two calibration procedures along with the sample sizes upon which they were based. In addition, Figures A-1, A-2, A-3, and A-4 in Appendix A present histograms of the distributions of the item parameter estimates.

Table 1  
Descriptive Statistics of Item Parameter  
Estimates for the Two Models

	One Parameter Model	Three Parameter Model		
	E <sub>i</sub>	a <sub>i</sub>	b <sub>i</sub>	c <sub>i</sub>
Mean	- .172	.990	- .519	.121
S.D.	1.467	.533	1.529	.042
Low	-2.821	.118	-3.624	.023
High	3.559	2.000 <sup>a</sup>	5.952	.270
Sample Size	1,000	1,541	1,541	1,541
No. of Items	72	72	72	72

<sup>a</sup> The LOGIST program imposes the restriction that discrimination estimates must stay in the range from .01 to 2.00.

In general these item parameters fall within the expected range for both models, and none of the items were deleted from the pools due to unusual parameter estimates. As can be seen in Table 1, the average difficulty (b) for the items in the three-parameter calibration is  $-.519$ , indicating that the calibration sample found the vocabulary items to be relatively easy to answer correctly. However, the distribution of these values shown in Figure A-2 is clearly peaked, rather than exhibiting the preferred rectangular distribution (Urry, 1977). Although the average item discrimination (a) is approximately  $1.0$ , nearly half (32) of the items in the pool fail to meet Urry's criterion of  $.80$  as the minimum acceptable value for items. However, it will be seen later that, since items are selected for administration based on the information function, in actual practice items with discrimination values below  $.80$  will rarely be administered in the three-parameter tailored test. The average guessing parameter (c) is well below Urry's recommended maximum value of  $.30$ , probably because the LOGIST procedure imposes severe restrictions on changes in the guessing parameter away from the initial  $.22$  estimate. For the one-parameter calibration, the average easiness (E) value of  $-.172$  is close to zero, but again the distribution is peaked, as can be seen in Figure A-1.

#### Tailored Testing Procedure

Once the calibrated item pools have been established, the three requisite components of the tailored testing procedure include (a) an item selection routine, (b) an ability estimation technique, and (c) a stopping rule to terminate the test. Each of these components will now be described for both the one- and the three-parameter tailored testing procedures used in this study.

For the one-parameter procedure, items were selected for administration based on difficulty values ( $b_i$ ). The procedure began with an ability estimate of  $+.50$  for the examinees. For example, if a person's initial ability was randomly assigned to be  $+.50$ , the testing procedure searched the item pool to locate the first item encountered with difficulty value equal to that ability estimate, within a  $+.30$  acceptance range. If the examinee answered the first item correctly, the next item to be administered was the first item in the pool at a predetermined, fixed stepsize ( $.693$ ) away in a positive direction, still within a  $+.30$  acceptance range, i.e. a more difficult item. On the other hand, an incorrect response to the item resulted in the selection for administration of the first item that was  $-.693$  away, i.e. an easier item. The  $.693$  fixed stepsize value had been previously determined through an analysis of tailored testing operation, (Reckase, 1976), being simply the value of  $\ln 2$ .

When at least one item had been answered correctly and one answered incorrectly, the ability level of the examinee was estimated using an empirical maximum likelihood procedure. At this point, the next item to be administered was selected so that it had a probability of  $.50$  of being answered correctly for the estimated ability level. For the one-parameter model, this was an item with difficulty equal to the ability estimate. The first item encountered in the pool that was nearest to the desired difficulty value within the  $+.30$  acceptance range was thus administered. Ability estimation was then performed after each subsequent item was answered. An empirical maximum likelihood procedure was used in an

iterative search to determine the mode of the likelihood distribution, which became the new ability estimate (See Patience, 1977 and Reckase, 1974 for a more thorough discussion).

There were two stopping rules for the procedure. First, if no items existed in the pool that fell within the  $\pm .30$  acceptance range, then no further items could be administered which were appropriate. Alternatively, the procedure terminated when a maximum of 20 total items had been administered.

For the three-parameter procedure, items were selected for administration to maximize the value of the information function, which describes for each item the potential contribution to the estimation of the examinee's ability. It is important to realize that for the one-parameter model, selecting items on the basis of easiness was equivalent to selecting items to maximize the information that an item provided about a person's ability. That is, information was maximized for the one-parameter model when the item was of exactly appropriate difficulty for a given ability level (when the item difficulty equaled the estimated ability).

However, for the three-parameter model, the information function was more complex. In particular, the added discrimination and guessing parameters played a crucial role in determining the amplitude of the information curve. The formula used to compute item information for the three-parameter logistic model was given in Birnbaum (1968) as

$$I(\theta_j, u_{ij}) = D^2 a_i^2 \psi[DL_i(\theta_j)] - D^2 a_i P_{ij}(\theta_j) \psi[DL_i(\theta_j) - \log c_i] \quad (3)$$

where  $I(\theta_j, u_{ij})$  is the information of Item  $i$  at ability level  $\theta$  for Person  $j$ , given item response  $u_{ij}$ ;  $Li(\theta_j) = a_i(\theta_j - b_i)$ ;  $P_{ij}(\theta_j)$  is the probability of a correct response to Item  $i$  given ability level  $\theta_j$ ;  $\psi(x)$  is the logistic probability density function; and the other parameters have their meanings mentioned previously. The total test information was then simply the sum of the item information (Birnbaum, 1968) given by

$$I(\theta) = \sum_{i=1}^n I(\theta_j, u_{ij}). \quad (4)$$

The tailored testing procedure for the three-parameter logistic model began the same way as the one-parameter procedure already described. Namely, a fixed .693 stepsize was used to select items until at least one correct and one incorrect response were obtained. At that point, an ability estimate was computed, again using the maximum likelihood technique. However, to select the next item to be administered, the item pool was searched for the item which was most informative (i.e.  $I(\theta_j, u_{ij})$  was maximal) for that particular ability estimate. This process was repeated until one of the two stopping rules was encountered: either no item in the pool was available with  $I(\theta_j, u_{ij}) > .70$  or a total of 20 items had been administered.

A problem in using maximum likelihood ability estimation with the three-parameter model was that the procedure was sometimes unable to converge on an ability estimate. Usually the problem was that the items were too difficult for an examinee, so that a string of incorrect responses was obtained, causing the mode of the likelihood distribution to be at the guessing level. In such cases, no maximum of the likelihood function could

be calculated. The current study utilized an arbitrary procedure in an attempt to "fix" the nonconvergence problem, which was to give an ability estimate that was .693 less than the previous ability estimate in order to select the next item. As will be seen later, this attempt was only partially successful.

### Design

The study employed a counterbalanced design in which there were two separate test sessions one week apart for each examinee, with both the one- and three-parameter tests administered at each session. In addition, the display mode of the cathode ray tube (CRT) screens, i.e. white-on-black or black-on-white, was varied. The counterbalancing resulted from the reversal of the order of the test model used from one test session to the next, as well as alternating the mode of CRT screen display. For example, if an examinee took the test for the first session on the black-on-white screen display, then the second session would be on the white-on-black CRT screen, and vice versa. Lighting conditions in the room were held constant, as were the CRT screen brightness and contrast controls, once a suitable adjustment level was obtained through the agreement of six judges.

The tailored test was arranged so that the examinees could not have perceived receiving two tests during each session. The computer program began administering the second test immediately after arriving at an ability estimate from the first test, so there was no pause between them. Since both item pools were identical in content, however, the examinees were told that occasionally they would receive the same test item to answer twice. The tailored tests were administered on Applied Digital Systems (ADDs) Consul 980 cathode ray tube terminals which were connected to an IBM 370/168 computer through a timesharing system.

The basic purpose of the test-retest design was to facilitate comparisons between the two tailored testing procedures (one-parameter vs. three-parameter) on the basis of reliability, information value of the tests, convergence rates to ability estimates, lack of fit of the models to the data, and attitude changes over test sessions, to name just the primary analyses performed.

### Sample

The subjects taking part in the study were undergraduate and graduate students enrolled in educational psychology and measurement courses at the University of Missouri - Columbia. A total of 142 students took part in the study. However, there were 14 instances in which data were missing for either the first or the second test session. In addition, there were 39 total examinees for whom the three-parameter test procedure failed to converge at an ability estimate. Thus, complete data were obtained on a total of 128 subjects. For 89 of these cases, the three-parameter procedure converged properly. All students received extra credit points toward course grades for participation in the study.

### Attitude Survey

At the end of each tailored test session, all examinees were asked to respond to a 20 item attitude questionnaire. The questionnaire was intended to measure the attitudes of the examinees toward the tailored testing procedure on the dimensions of difficulty, anxiety, time pressure, and motivation. All 20 statements were written in Likert scale fashion with a five point scale of response alternatives after each statement. (A complete sample of the attitude scale is reproduced in Appendix B.) The items were scored on a scale from 1 to 5 with 1 signifying that the item response reflected an unfavorable attitude toward tailored testing and 5 indicating a favorable attitude.

### Analyses

The measure used to determine the goodness of fit of the observed data to the models was the mean squared deviation (MSD) statistic (Reckase, 1977). This statistic was calculated by squaring the differences for each person between the actual response to an item and the probability of a correct response predicted by the model. These squared differences were then summed across the response string and divided by the number of items administered. The formula used for the MSD statistic was:

$$MSD_j = \frac{\sum_{i=1}^N (u_{ij} - p_{ij})^2}{N} \quad (5)$$

where  $MSD_j$  = the mean squared deviation for person  $j$

$u_{ij}$  = the actual response to item  $i$  by person  $j$

$p_{ij}$  = the probability of a correct response to item  $i$  by person  $j$  predicted by the model.

$N$  = the number of items in the tailored test

Thus each examinee had two MSD values calculated, one for the one-parameter and one for the three-parameter models. Theoretically the MSD statistic had a possible range of from 0 to 1. A value of 0 was obtained if the model fit the item responses perfectly and 1 was obtained when there was total lack of fit. In actual practice, however, the value of the MSD for an examinee rarely exceeded .25 for either model, the value obtained when all of the ICC's are flat.

A systematic sample of 22 examinees was taken to compare the two models using the MSD statistic. A t-test was used to analyze these data. It was desired that MSD values be computed across the whole range of ability estimates yielded in the tailored tests, so that sampling was systematic instead of random to insure this representative coverage of ability. Although the complete sampling distribution of the MSD statistic was unknown, it was deemed justifiable to use the t-test to compare the MSD results for the two models, since previous research had shown the MSD distribution to be approximately normal (Reckase, 1977).

The reliability comparison of the two models was computed by correlating the ability estimates from the two sessions. These coefficients were

not test-retest reliability measures, but rather were hybrids of test-retest and equivalent forms reliability. Since a different entry point into the item pool was used for each session (either +.5 or -.5) and since changes in response strings resulted in different paths through the item pool, it was impossible for an examinee to receive exactly the same tailored test twice. However, numerous items were repeated over sessions as a function of the consistency in ability estimation for a person since items were selected from the same pool. Thus the reliability measure was neither a true test-retest, nor a true equivalent forms, but rather a mix between the two. The reliabilities were compared using a t-test based on the usual  $r$  to  $z$  transformation.

Correlation analyses were conducted among a great many variables measured by the study, but two were of major interest. First, the correlations of the ability estimates yielded by the one- and three-parameter models from the two testing sessions were compared. Second, the correlation was calculated between the ability estimates and the outside criteria of performance, namely, traditional paper-and-pencil exam scores over the students' course material. The course exams either consisted of a series of three 50 item multiple choice tests which covered introductory measurement and evaluation concepts, or else one 35 item multiple choice final exam over introductory educational psychology material, depending on the course in which the student was enrolled. The purpose of these correlations was to determine the degree to which the two procedures of the models were measuring the same thing, and whether one model performed better than the other in the prediction of the outside criteria.

Numerous descriptive statistics were compiled for the one- and three-parameter tailored tests which included such variables as average test length, average test difficulty, percentage of test items in common over the two sessions, etc. Comparisons of the two models were made on the basis of these data. Where significant differences were found, the effects on reliability were partialled out.

Information analyses were performed to compare the two models in terms of relative efficiency (Birnbaum, 1968), the ratio of information provided by each model's tailored test to the information provided by a thirty item paper-and-pencil test. In addition, plots were drawn of the information functions for the tailored and paper-and-pencil test, as well as plots of information on a per item basis. Again, the plots were constructed with cases from a systematic rather than a random sample to insure that data across the whole range of ability estimates yielded by the tailored tests was available.

Convergence plots were also drawn for the tailored tests taken by each examinee over both sessions. On one axis were plotted the ability estimates calculated after each item was administered, and on the other axis were plotted the trials (items received). The purpose of the convergence plots was to provide a graphic representation of the rates at which the one- and three-parameter tailored tests converged to stable ability estimates. Although the fact of the ability estimates having different scales prohibited direct comparisons of the two models, still it was useful to make subjective judgments through careful scrutiny of the convergence plots. Several representative plots were, therefore, selected for discussion in this report.

A final set of analyses were conducted on the attitudinal data collected from the Likert questionnaire administered after each session. The initial effort was to determine the various characteristics of the attitude scale itself, followed by efforts to analyze the attitude data for tailored testing implications. Two different types of factor analysis procedures were tried out on the scale to determine its dimensionality, including a principal components with varimax rotation, as well as orthogonal procrustes rotation. The latter technique was run to compare the factor structure from the first administration of the questionnaire to the second.

Once the factor structure was determined, individual items on the scale that loaded together were examined to label the factors. Coefficient alpha reliabilities were then calculated for each factor, as well as for the total scale. Traditional item analysis correlations were also used to determine the contribution of each item on the scale to the total score obtained. And finally, factor scores were generated for the examinees for subsequent analysis.

Frequencies of responses to the five scale points for each item were determined for both test sessions, to provide an indication of agreement or disagreement with the various statements about tailored testing. In addition, several multivariate analyses of variance (MANOVA's) were performed on the attitude data. One attempted to determine the differences in attitude from the first test session to the second, another from one screen display format to the other. Both of these comparisons were made on the basis of factor scores as well as item raw scores.

## Results

### Goodness of Fit

The results of the MSD statistic used to compare the goodness of fit of the one- and three-parameter logistic models are presented in Table 2. The values of MSD for 22 cases for each model are shown, along with the means, standard deviations, and the results of a dependent t-test analysis of the data. As has been mentioned earlier, the 22 cases were obtained from a systematic sample of 89 examinees to insure that data were available across the whole range of ability estimates yielded by the tailored tests. The results of the t-test showed that the MSD statistic was significantly smaller for the three-parameter ( $p < .05$ ) indicating better fit of the model to the observed responses.

### Reliability

The correlation matrix in Table 3 reports the coefficients obtained from intercorrelating the ability estimates yielded by the two models in the tailored testing study. Of special interest are the correlations between the ability estimates from the first one-parameter logistic tailored test (1PL 1) and the second one-parameter tailored test (1PL 2), and the first three-parameter test (3PL 1) with the second three-parameter test (3PL 2). The .61 correlation value shown in Table 3 is the reliability coefficient

Table 2

Goodness of Fit Comparison  
Using the MSD Statistic

Observations	One Parameter MSD	Three Parameter MSD
1	.198	.184
2	.197	.206
3	.212	.158
4	.214	.100
5	.083	.143
6	.203	.098
7	.202	.208
8	.187	.156
9	.208	.153
10	.204	.140
11	.192	.171
12	.083	.133
13	.215	.267
14	.196	.191
15	.164	.198
16	.194	.144
17	.203	.166
18	.203	.126
19	.183	.247
20	.214	.149
21	.182	.022
22	.188	.185
$\bar{x}$	.188	.161
$s_x$	.055	.063

$t_{(21)} = 2.086$  ( $p < .05$ )

Table 3

Ability Estimate Correlations

Variables	1	2	3	4	5	6	7	8
1. 1PL 1		.61(.55) <sup>a</sup>	.96	.53	.57	.58	.53	.59
2. 1PL 2			.53	.90	.68	.70	.63	.69
3. 1PLEQI 1				.47	.49	.53	.44	.55
4. 1PLEQI 2					.52	.51	.47	.49
5. 3PL 1						.77(.36) <sup>a</sup>	.90	.76
6. 3PL 2							.79	.96
7. 3PLEQI 1								.78
8. 3PLEQI 2								

( )<sup>a</sup> indicates the inclusion of 39 cases for which the three-parameter test did not converge at an ability estimate. All other correlations are based on 89 cases.

for the one-parameter tailored test. This is to be contrasted with the .77 reliability coefficient obtained from the first three-parameter tailored test. The difference between these two reliabilities is statistically significant ( $p < .05$ ).

It is very important to note, however, that these reliabilities are based on only 89 rather than 128 cases. The values in parentheses in Table 3 reflect the reliability coefficients when data from 128 cases were included. The differences are due to the failure of the three-parameter logistic tailored test to converge at ability estimates for 39 cases, about one-third of the total. The nonconvergence of the three-parameter procedure will be discussed later, but notice the dramatic effect that including the 39 nonconvergence cases had on the reliability of the three-parameter tailored test (from .77 to .36). The one-parameter reliability also dropped slightly from .61 to .55. This decrease might have been due to a restriction in the range of ability estimates. However, the difference between the reliabilities for 128 cases (.36 vs. .55) was not statistically significant. Notice in this regard that although the absolute difference of .19 is greater than the absolute difference of .16 between the reliabilities for 128 and 89 cases respectively, the former value is not significant due to the nonlinearity of Fisher's  $r$  to  $z$  transformation used in the significance test.

At this point, it seems appropriate to discuss the rationale for what might appear to be an arbitrary decision to drop the 39 nonconvergence cases from the reliability analysis. The primary reason was that subsequent tailored testing research by the authors, which also used the three-parameter logistic model in conjunction with maximum likelihood ability estimation, showed the nonconvergence problem to be very minimal. In this latter case, the item pool was of appropriate difficulty for the examinees (i.e. the items covered their course material and had been previously calibrated using the same ability level of students) and the nonconvergence of ability estimation occurred in only nine out of 110 tailored tests. Therefore, the hypothesis was that the high rate of nonconvergence was due to the item pool being too difficult for the examinees, resulting in high guessing. For example, the mean one-parameter ability estimate for the 39 nonconvergence cases was only  $-.80$  compared to a mean of  $0.00$  for the calibration sample (see Table 4). For this reason, we decided to disregard the 39 nonconvergence cases for the current study in some of the data analyses and considered them to be a temporary source of poor ability estimation for the three-parameter tailored testing procedure.

Since it was common for each tailored test administered to an examinee to have a different number of test items, and since test length often substantially affects the reliability of a test, another comparison was undertaken in which all the ability estimates were equated for test length. The correlation between the first and second one-parameter tailored test ability estimates, .61, was compared to the correlation between the first and second three-parameter ability estimates for tests with an equal number of items presented (3PLEQ1 vs. 3PLEQ2). The resulting difference between these correlations, .61 and .78 was still significant ( $p < .05$ ).

Another factor was investigated for possible confounding of the reliability coefficients, namely, the number of test items in common from one

test to another. It was found, for instance, that the mean percentage of items in common between tailored tests one and two for the three-parameter model was 85%. For the one-parameter tailored tests, this value averaged only 20%. The reason for this substantial difference in the number of items in common over test sessions for the two models was related to their respective item selection procedures. Since the three-parameter procedure selected items on the basis of the highest information value for an ability estimate, and since items with moderately high discrimination values yield good information over a relatively broad ability range, there was a tendency for the same items to be administered to an examinee at both test sessions. For the one-parameter model, however, where the items administered were the first ones encountered within an acceptance range about the ability estimate, different items would be selected depending on the ability estimate, ignoring the item discrimination. Thus, there was a much lower chance that an examinee would receive the same items from one test to the next for the one-parameter test.

In order to determine the effect of proportion of items in common between tests on reliability, two approaches were used. First, correlation coefficients were computed between the absolute difference in the two ability estimates for the one-parameter tests and the corresponding proportion of test items in common for each examinee. The same correlation was figured for the three-parameter tests. For the one-parameter procedures,  $r$  was found to be  $-.634$ , while  $r$  was  $-.496$  for the three-parameter case (the difference between these correlations was not significant). Thus examinees with small differences between their first and second ability estimates for both models tended to have tailored tests with high proportions of items in common, and vice versa. This finding implied that number of items in common had some effect on reliability, but the problem was to determine its magnitude.

To accomplish this partial correlation coefficients were computed using the formula given below.

$$r_{12 \cdot 3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \quad (6)$$

Depending on whether the partial correlation coefficient was being calculated for the one- or the three-parameter tailored tests,  $r_{12}$  was either the correlation between the two one-parameter ability estimates or the two three-parameter ability estimates. The values for  $r_{13}$  and  $r_{23}$  were the correlations between each of the respective ability estimates and the third variable, namely the proportion of common test items. For the one-parameter test,  $r_{13} = .064$  and  $r_{23} = .111$ ; while for the three parameter test, these values were  $r_{13} = .395$  and  $r_{23} = .211$ . However, when these values were substituted into formula 6 in order to adjust the original reliability  $r_{12}$ , the reduction due to proportion of items in common was negligible (less than .01). This result indicated that although the proportion of common items across tests differed markedly for the one-parameter compared to the three-parameter test, the proportion of items in common did not materially contribute to either reliability value.

Table 4 represents several additional descriptive statistics computed for the one- and three-parameter logistic tailored tests. For example, the mean test difficulty for both procedures was about the same, close to .50, showing that both procedures were administering items of appropriate difficulty for 89 of the examinees. Lord (1970) has indicated that "measurement is most effective when the examinee knows the answers to only about half of the test items (p. 140)." In addition, the three-parameter test was slightly longer on the average than the one-parameter test.

Table 4  
Descriptive Statistics

Variable	One Parameter Tailored Test	Three Parameter Tailored Test
Mean # of items administered	15.07 (13.13)	19.39 (9.60)
Mean # of items correct	7.45 (5.39)	8.95 (2.35)
Mean test difficulty	.49 (.41)	.49 (.24)
Mean ability estimates	.44 (-.80)	-.77 (-5.48)

n = 89

Note: Values in parentheses are for the 39 nonconvergence cases.

Figures C-1 through C-4 in Appendix C, however, show that the distributions of ability estimates obtained from the tailored tests are all approximately normal, with only minor variations from the first to the second session.

#### Other Correlation Analyses

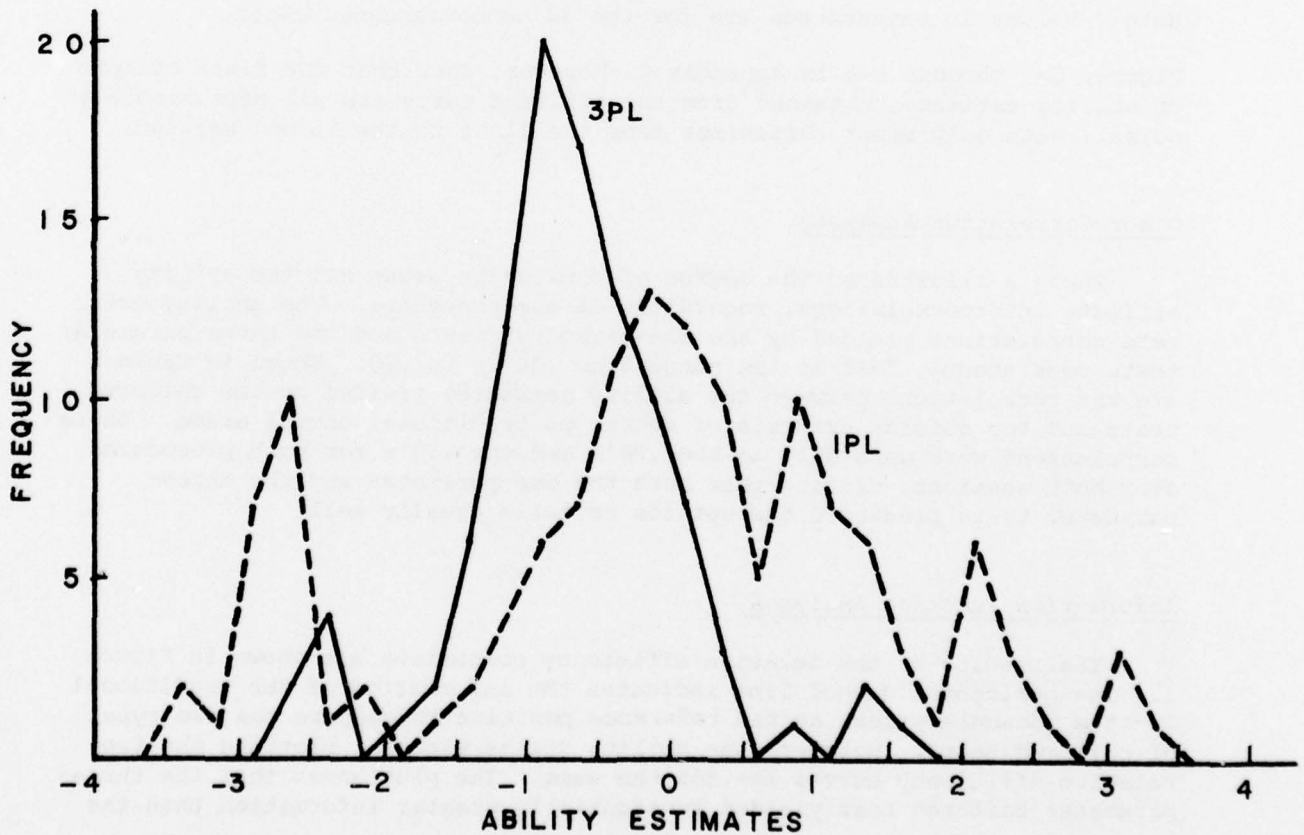
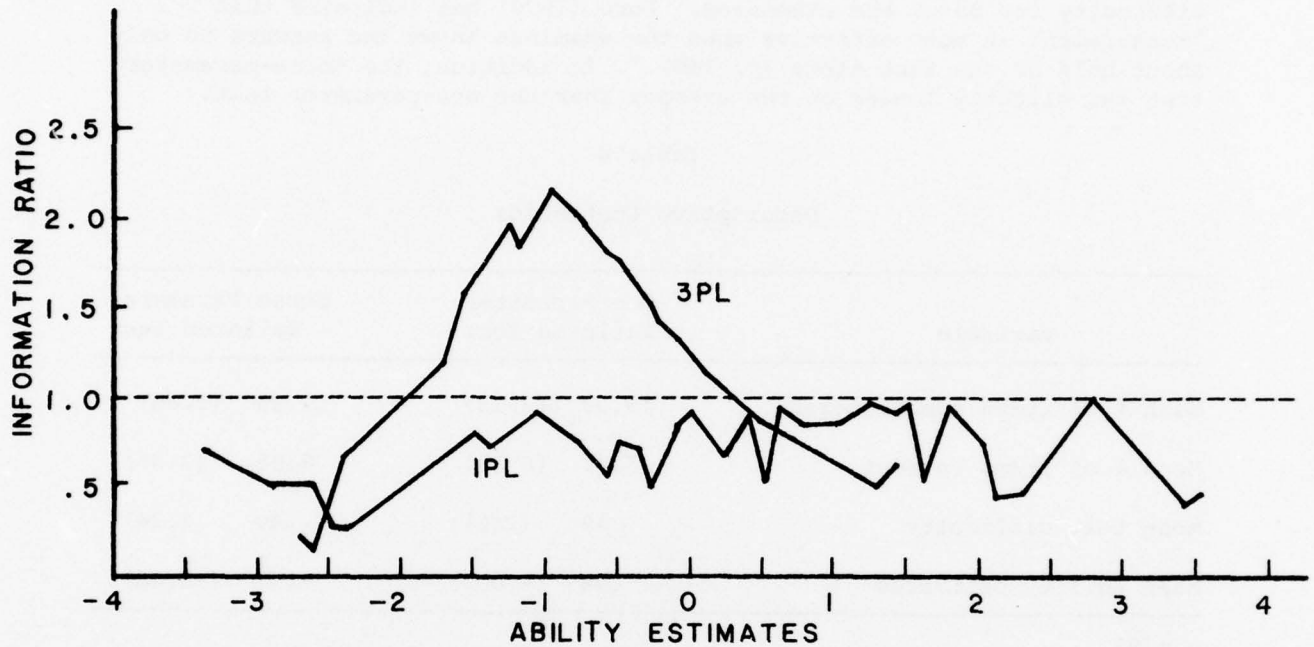
Table 3 illustrates the degree of similarity among all the ability estimate intercorrelations, regardless of the procedure. The ability estimate correlations yielded by the one-parameter tests and the three-parameter tests consistently fall in the range from .44 up to .70. Shown in Table 5 are the correlations between the ability estimates yielded by the tailored tests and the outside criteria of scores on traditional course exams. These correlations were generally in the .20's and the .30's for both procedures over both sessions, meaning that both the one-parameter and the three-parameter tests predicted the outside criteria equally well.

#### Information Function Analyses

The results of the relative efficiency comparison are shown in Figure 1. The horizontal dashed line indicates the information of the traditional 30-item vocabulary test as the reference position to compare the two types of tailored tests. However, the ability scales used for plotting the two relative efficiency curves are not the same. The plot shows that the three-parameter tailored test yielded substantially greater information than the

FIGURE 1

RELATIVE EFFICIENCY



Notes: Nonconvergence cases deleted.

Table 5  
Correlations of Ability Estimates  
With Outside Criteria

Variables	1PL1	1PL2	3PL1	3PL2
1. 1st measurement course exam score	.36(49)	.29(49)	.24(49)	.09(49)
2. 2nd measurement course exam score	.21(49)	.33(49)	.25(49)	.27(49)
3. 3rd measurement course exam score	.32(49)	.19(49)	.32(49)	.30(49)
4. Final exam score for educ. psych. course	-.00(22)	.18(21)	.42(22)	.32(21)

Note: The values in parentheses indicate the number of cases upon which the correlations are based.

traditional test, but only in a peaked fashion for ability estimate levels between -2.0 and +.50, falling off sharply outside this range. However, at no point did the one-parameter tailored test exceed the traditional test information, and its information curve was rectangular rather than peaked. Also shown in Figure 1 are the frequency distributions of ability estimates obtained from the two procedures. Note that the information from the three-parameter test is greatest where most of the examinees were concentrated.

#### Convergence Plots

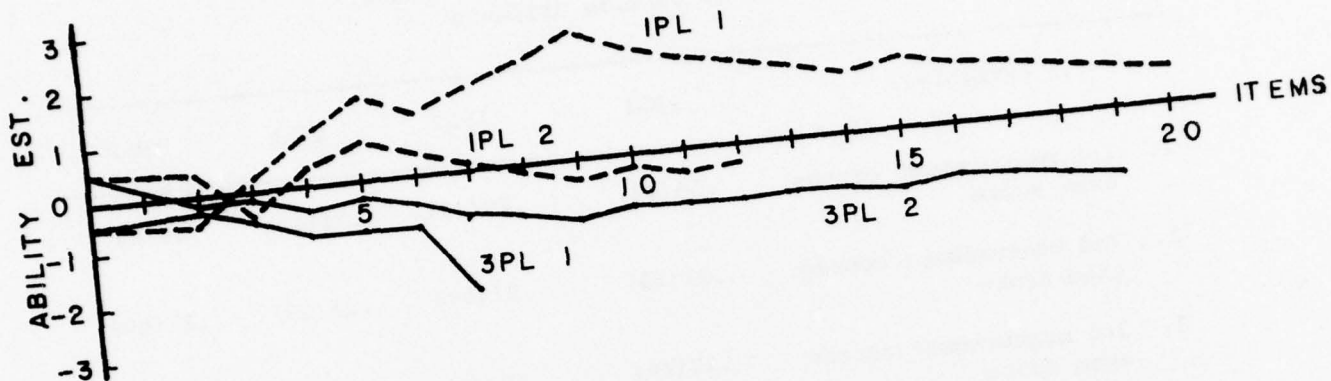
In Figure 2 are pictured four individual tailored testing convergence plots, including good and poor examples of convergence for each of the two types of tailored tests. Plot 2-A shows a case where neither procedure converged very well, 2-B a case where the one-parameter test did well but not the three-parameter test, 2-C a case in which the three-parameter test converged better than the one-parameter test, and 2-D where both procedures converged nicely. A subjective classification method applied to 44 separate cases resulted in the following breakdown: 2-A, 7 plots; 2-B, 5 plots; 2-C, 18 plots; and 2-D, 14 plots. However, recall that in 39 cases, not included in the above categories, the three-parameter tailored testing procedure failed to converge at all.

#### Attitude Scale Characteristics

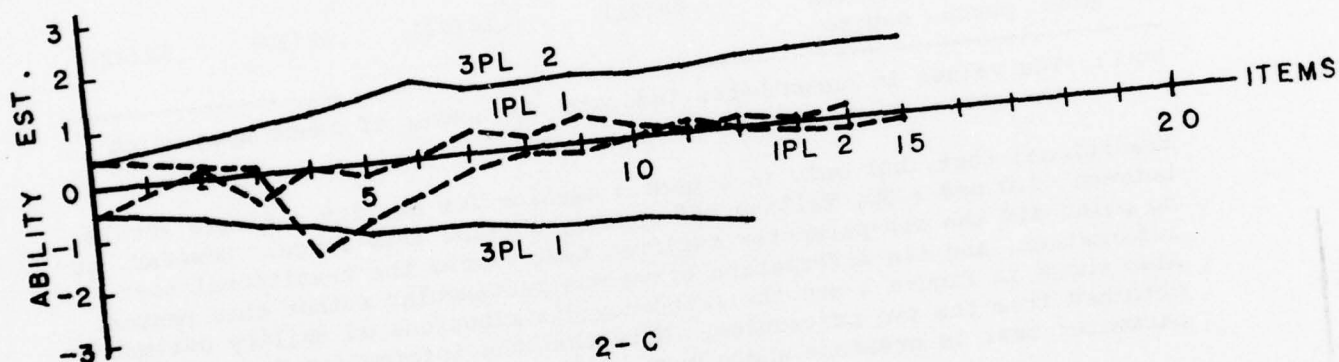
The varimax rotated factor loading matrix that was obtained from a principal components analysis of the first administration of the attitude scale is shown in Table 6. There were six factors present with eigenvalues greater than one, which accounted for 62% of the variance. The underlined

-16-  
FIGURE 2  
CONVERGENCE PLOTS

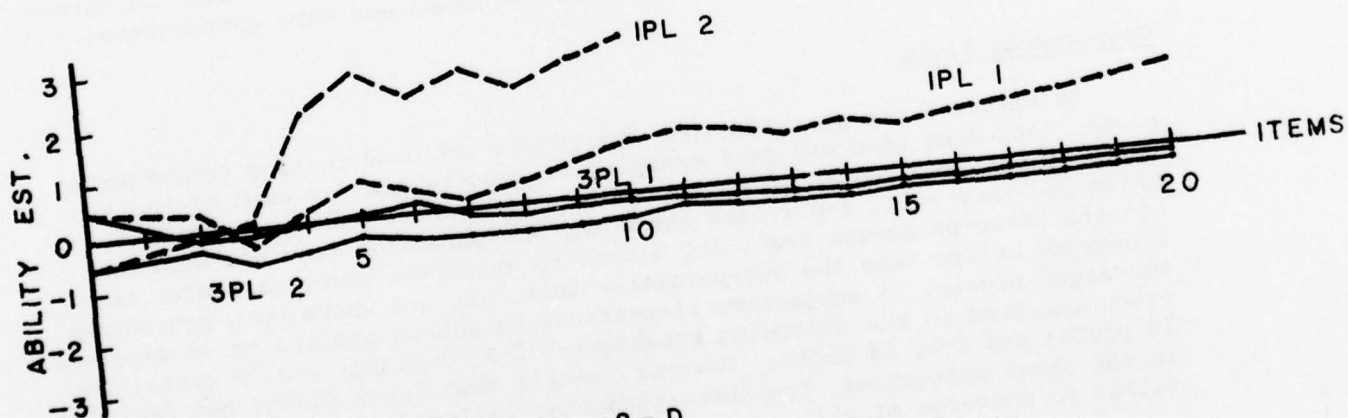
2-A



2-B



2-C



2-D

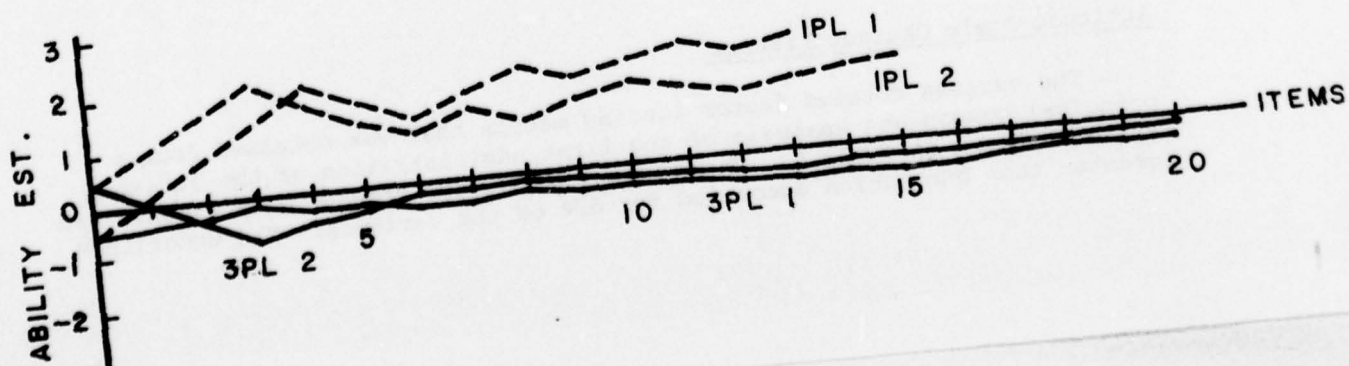


Table 6

Rotated Factor Pattern for First  
Attitude Scale Administration

Item No.	I	II	III	IV	V	VI
1	<u>.68</u>	.01	-.19	.18	-.06	.18
2	.08	-.08	.13	.05	-.01	<u>-.82</u>
3	-.10	-.21	-.16	.11	<u>-.63</u>	<u>-.34</u>
4	<u>.45</u>	.09	.20	.25	-.05	<u>-.52</u>
5	<u>-.56</u>	-.17	.43	.08	-.03	-.19
6	.17	-.01	.20	-.11	<u>-.74</u>	.20
7	.00	<u>-.83</u>	.10	.06	.03	-.12
8	.01	.12	.28	<u>.74</u>	-.23	-.08
9	.07	.04	.13	.07	<u>.67</u>	.14
10	<u>.68</u>	-.09	.21	-.16	-.06	-.10
11	.08	-.14	.13	<u>.63</u>	.27	.14
12	.08	<u>-.71</u>	.04	-.02	-.36	.16
13	<u>.67</u>	<u>-.13</u>	.24	.13	-.11	-.25
14	<u>.15</u>	-.28	-.12	.04	<u>-.76</u>	.13
15	<u>.76</u>	-.06	.08	.03	.05	-.14
16	.03	-.06	-.13	<u>.75</u>	.04	-.24
17	.06	<u>-.88</u>	-.02	.01	-.17	-.08
18	-.00	.06	<u>.77</u>	.07	.13	.11
19	.11	-.07	<u>.71</u>	-.04	-.08	-.14
20	.09	-.07	<u>.58</u>	.17	.18	-.20

Note: The underlined values indicate the highest loading of an item on a factor. Broken underlines indicate other high loadings.

values in the table indicate the highest factor loading for each item on the scale among the six factors. A subjective examination of the items loading on each factor resulted in the following factor labels:

- factor I - anxiety
- factor II - test satisfaction
- factor III - motivation
- factor IV - cathode ray tube (CRT) characteristics
- factor V - perceived difficulty
- factor VI - time pressure

Since the examinees responded to the attitude scale after each test session, data were also available from the second administration of the attitude scale. Again a principal components analysis with a varimax rotation was conducted, and again six factors were present with eigenvalues greater than one, with 63% of the variance accounted for. The rotated factor loading matrix is presented in Table 7. Many of the factors changed in order, but the same six factors were identified. A few of the items on the attitude scale switched factors, but in general the pattern of components was the same.

The labeled factors are listed below:

factor I - perceived difficulty  
 factor II - CRT characteristics  
 factor III - motivation  
 factor IV - anxiety  
 factor V - test satisfaction  
 factor VI - time pressure

Table 7

Rotated Factor Pattern for Second  
 Attitude Scale Administration

Item No.	I	II	III	IV	V	VI
1	-.13	-.02	-.15	<u>.79</u>	-.07	-.08
2	-.11	.26	-.04	<u>.19</u>	<u>.51</u>	.10
3	<u>.61</u>	.03	-.08	-.11	.01	-.30
4	.05	.37	.16	.26	-.00	<u>.51</u>
5	.15	.13	.41	<u>-.68</u>	-.08	-.02
6	<u>.81</u>	-.12	.10	<u>.10</u>	-.06	.05
7	.19	-.13	.10	-.12	<u>.85</u>	-.03
8	.11	<u>.85</u>	.03	-.12	.03	.10
9	<u>-.69</u>	-.09	.13	.04	-.11	-.06
10	.13	.22	.13	<u>.71</u>	-.02	.07
11	-.03	.06	-.07	-.13	.00	<u>.87</u>
12	<u>.67</u>	.02	-.12	-.03	.36	.03
13	.15	<u>.53</u>	.24	<u>.45</u>	.01	-.11
14	<u>.81</u>	.10	.06	-.02	.12	.05
15	-.01	<u>.48</u>	-.08	<u>.38</u>	-.12	.09
16	-.05	<u>.82</u>	.05	.03	.09	.08
17	.34	.00	.05	-.11	<u>.79</u>	-.06
18	-.25	-.00	<u>.73</u>	-.04	.03	-.02
19	.11	-.04	<u>.80</u>	-.06	.00	.14
20	-.03	.16	<u>.80</u>	-.07	.06	-.09

Note: The underlined values indicate the highest loading of an item on a factor. Broken underlines indicate other high loadings.

In order to obtain an indication of the reliability of the attitude scale, several techniques were utilized. First, an orthogonal procrustes factor analysis was run on the data from the two scale administrations. Given the rotated factor loading matrix results of the first administration to be used as the target or goal matrix, this procedure attempted to rotate the factor matrix from the second administration to be equal to the target matrix plus error, by postmultiplication of the second matrix by a transformation matrix. In other words, the program solved the equation  $AT = B + E$ , where B was the target matrix, A was the second session factor matrix, E was a matrix of residuals, and T was a transformation matrix. By comparing the AT matrix to B, as well as observing the sizes of the residuals in E, it was possible to estimate the degree to which the two attitude scale administrations measured the same things. In general the results showed the scale to

have good reliability. The largest residual obtained was .49, although the majority of residuals were less than or equal to .10, with most of the larger residuals found on the weakest factor. Table 8 shows the B target matrix, Table 9 shows the AT matrix, and Table 10 contains the residuals. The slight differences between the values in Table 6 and those in Table 8 are due simply to the use of two different packages, SAS and SOUPAC respectively, to perform the principal components analyses.

The second reliability measure of the attitude scale was a traditional test-retest reliability coefficient. The value obtained was  $r = .57$  based on 130 cases with data for both sessions. A total attitude score for each examinee was obtained on the scale by summation of the scores on the 20 individual items. The reliability was then simply the correlation between the sets of total attitude scores for the two tailored testing sessions, assuming that no attitude changes had taken place.

In addition, discrimination indexes were calculated for all of the items on the scale by correlating individual item scores with total scores for each examinee. The results over the two test sessions are shown in Table 11. Most of the discrimination values are fairly high, being in the range of the +.30's and up, despite the fact that the attitude scale comprised six factors as previously discussed.

Finally, coefficient alpha reliabilities were calculated for the factors of the scale as well as for the total scale itself. The results are shown in Table 12. In general the reliability coefficients for the factors are fairly high, with the only exception being the anxiety factor. The coefficient alphas for the total attitude scale for the two sessions, .67 and .57 respectively, are quite comparable to the test-retest reliability value of .57 that was previously reported.

### Attitude Scale Results

Having discussed the psychometric characteristics of the attitude scale, we turn now to summarizing the results obtained from the administration of the scale. In Tables 13-18 are listed the response percentages for the various factors measured by the attitude scale for the two tailored testing sessions. In general there appear to be several apparent contradictions of responses on the various items within a factor or subscale. For instance in Table 13 we see that the examinees were not nervous about the prospect of taking the test on the computer terminal but did experience stress during the test itself. In Table 14, the response percentages indicate that the examinees felt many of the test items to be too easy, yet they were not confident of having performed well. Table 15 responses show that motivation levels were high during the tailored test, which may have heightened the perceived time pressure responses reflected in Table 18. The examinees felt considerable time pressure in spite of the fact that they had as much time to respond to each item as desired. The response data in Table 16 reveals that the CRT screen did not cause eye discomfort, but that the items were difficult to read. And finally, in Table 17, the examinees felt that the tailored test did a good job of measuring their vocabulary ability, but that it did not reflect their "true" vocabulary knowledge. Some of these apparent anomalies will be discussed later in this report.

Table 8  
Target Matrix B

Item No.	I	II	III	IV	V	VI
1	<u>.70</u>	.04	-.05	-.16	.11	-.18
2	<u>.08</u>	.04	.04	.06	.06	<u>.84</u>
3	-.06	<u>.67</u>	.09	-.22	.11	.27
4	<u>.46</u>	.00	-.08	.21	.24	<u>.47</u>
5	<u>-.58</u>	.02	.24	.34	.09	.27
6	<u>.18</u>	<u>.72</u>	.02	.21	-.13	-.20
7	.03	-.06	<u>.82</u>	.09	.06	.06
8	.06	.23	-.10	.32	<u>.71</u>	.07
9	.11	-.66	-.06	.15	<u>.04</u>	-.07
10	<u>.67</u>	<u>.05</u>	.14	.21	-.17	.14
11	<u>.06</u>	-.25	.15	.14	<u>.65</u>	-.08
12	.02	.36	<u>.72</u>	.01	<u>.02</u>	-.06
13	<u>.65</u>	.08	<u>.19</u>	.25	.16	.28
14	<u>.17</u>	<u>.76</u>	.23	-.07	-.00	-.12
15	<u>.75</u>	-.01	.03	.10	.02	.15
16	<u>-.03</u>	-.02	.05	-.19	<u>.73</u>	.23
17	.02	.15	<u>.89</u>	-.06	<u>.02</u>	.06
18	.05	-.19	-.10	<u>.78</u>	.09	-.07
19	.07	.02	.08	<u>.73</u>	.02	.12
20	.05	-.17	.10	<u>.47</u>	.09	.38

Table 9  
Procrustean Rotated Pattern Matrix AT

Item No.	I	II	III	IV	V	VI
1	<u>.71</u>	-.07	-.19	-.24	-.29	.02
2	<u>.28</u>	-.22	<u>.34</u>	-.03	.16	<u>.35</u>
3	-.01	<u>.62</u>	.14	.04	-.03	-.18
4	<u>.45</u>	-.06	.08	.27	<u>.41</u>	.10
5	<u>-.49</u>	.12	.18	<u>.54</u>	<u>.32</u>	-.10
6	<u>.14</u>	<u>.74</u>	.15	.10	-.12	-.20
7	-.17	<u>.00</u>	<u>.81</u>	-.04	-.15	.30
8	.05	.26	-.10	.02	<u>.69</u>	.48
9	.07	-.62	-.22	.11	-.04	-.20
10	<u>.74</u>	.11	-.06	.14	-.00	.11
11	.11	-.25	.25	.12	<u>.58</u>	-.40
12	.06	.43	<u>.66</u>	-.05	.07	-.07
13	<u>.52</u>	.25	-.03	.17	.16	.43
14	<u>.06</u>	<u>.77</u>	.29	.06	.10	-.08
15	<u>.57</u>	-.03	.00	.01	.40	-.02
16	<u>.08</u>	.16	-.21	-.05	<u>.55</u>	<u>.61</u>
17	-.09	.17	<u>.82</u>	-.04	-.06	.25
18	-.11	-.27	-.05	<u>.63</u>	-.15	.33
19	-.16	.05	.06	<u>.70</u>	-.13	.28
20	.05	-.01	.05	<u>.80</u>	-.00	.22

Note: The underlined values indicate the highest loading of an item on a factor. Broken underlines indicate other high loadings.

Table 10

Residual Matrix E

Item No.	I	II	III	IV	V	VI
1	-.01	.11	.14	.08	.41	-.20
2	-.20	.26	-.30	.10	-.10	.49
3	-.05	.05	-.05	-.26	.13	.45
4	.01	.06	-.16	-.07	-.18	.37
5	-.09	-.10	.06	-.19	-.23	.37
6	.04	-.02	-.13	.11	-.00	.00
7	.21	-.06	.01	.13	.22	-.24
8	.02	-.03	-.01	.31	.02	-.41
9	.04	-.03	.15	.04	.09	.13
10	-.08	-.06	.20	.07	-.17	.03
11	-.05	.00	-.10	.02	.07	.33
12	-.04	-.06	.06	.06	-.05	.01
13	.13	-.17	.22	.07	-.01	-.15
14	.11	-.01	-.06	-.13	-.11	-.04
15	.18	.01	.03	.08	-.38	.17
16	-.10	-.18	.26	-.14	.18	-.38
17	.11	-.02	.08	-.00	.08	-.19
18	.16	.08	-.05	.15	.23	-.40
19	.23	-.02	.02	.03	.15	-.16
20	.01	-.16	.05	-.33	.09	.16

Table 11

Discrimination Indexes for Attitude Scale  
Items for Two Test Sessions

Item No.	Session 1	Session 2
1	.33	.08
2	.43	.38
3	.32	.31
4	.50	.43
5	.18	.25
6	.28	.31
7	.50	.32
8	.47	.55
9	-.02	-.20
10	.41	.40
11	.35	.24
12	.46	.44
13	.59	.47
14	.36	.47
15	.47	.36
16	.35	.49
17	.49	.43
18	.27	.24
19	.39	.34
20	.37	.42

Table 12  
Coefficient Alpha Reliabilities for the  
Attitude Scale Factors and Total Scale

Factor Labels	Items	Coeff. $\alpha$ Session 1	Coeff. $\alpha$ Session 2
Anxiety	1, 5, 10, 13, 15	.32	.21
Perceived Difficulty	3, 6, 9, 12, 14	.31	.46
Motivation	18, 19, 20	.79	.55
CRT Characteristics	8, 11, 16	.51	.53
Test Satisfaction	7, 17	.63	.96
Time Pressure	2, 4	.44	.49
Total Scale	all 20 items	.67	.57

Table 13  
Response Percentages for the Anxiety Factor  
for Items and Alternatives Over Both Sessions

1. During the test I was worried about how well I was doing.	<u>session 1</u>	<u>session 2</u>
strongly agree	9	2
agree	39	24
neutral	18	22
disagree	22	32
strongly disagree	12	20
5. I didn't care very much about how well I did on the test.	<u>session 1</u>	<u>session 2</u>
strongly disagree	2	4
disagree	17	18
neutral	30	26
agree	42	46
strongly agree	9	6
10. I was nervous about coming here to take this test.	<u>session 1</u>	<u>session 2</u>
strongly agree	0	1
agree	6	2
neutral	11	4
disagree	43	60
strongly disagree	40	33
13. The computer terminal made me nervous.	<u>session 1</u>	<u>session 2</u>
strongly agree	0	0
agree	3	1
neutral	6	4
disagree	58	68
strongly disagree	33	27
15. I felt considerable stress while taking the test.	<u>session 1</u>	<u>session 2</u>
strongly disagree	0	3
disagree	6	3
neutral	9	9
agree	54	63
strongly agree	31	22

Table 14

Response Percentages for the Perceived Difficulty  
Factor for Items and Alternatives Over Both Sessions

<hr/>		
3. I felt that many of the items were too difficult for me.	<u>session 1</u>	<u>session 2</u>
strongly disagree	24	14
disagree	51	53
neutral	15	18
agree	9	12
strongly agree	1	3
6. I think I did well on the test compared to other people.	<u>session 1</u>	<u>session 2</u>
strongly agree	6	3
agree	35	38
neutral	49	48
disagree	10	11
strongly disagree	0	0
9. I felt that many of the items on the test were too easy.	<u>session 1</u>	<u>session 2</u>
strongly disagree	1	1
disagree	2	1
neutral	8	9
agree	54	65
strongly agree	35	24
12. I feel that I did as well on this test as on other vocabulary tests I've taken.	<u>session 1</u>	<u>session 2</u>
strongly agree	17	11
agree	44	52
neutral	18	18
disagree	20	19
strongly disagree	1	0
14. I felt confident that I did well on the test.	<u>session 1</u>	<u>session 2</u>
strongly disagree	12	9
disagree	57	56
neutral	27	29
agree	4	6
strongly agree	0	0
<hr/>		

Table 15  
Response Percentages for the Motivation Factor  
for Items and Alternatives Over Both Sessions

18. I think I could have done better on the test if I had tried harder.	<u>session 1</u>	<u>session 2</u>
strongly disagree	0	1
disagree	7	11
neutral	14	12
agree	70	69
strongly agree	9	7
19. I was careful to try to select the best answer to each question.	<u>session 1</u>	<u>session 2</u>
strongly disagree	0	1
disagree	2	4
neutral	7	9
agree	69	72
strongly agree	22	14
20. I tried to finish the test quickly just to receive my 5 points credit.	<u>session 1</u>	<u>session 2</u>
strongly agree	0	0
agree	1	2
neutral	6	9
disagree	63	65
strongly disagree	30	24

Table 16  
Response Percentages for the CRT Characteristics Factor  
for Items and Alternatives Over Both Sessions

8. My eyes were uncomfortable when viewing the screen.	<u>session 1</u>	<u>session 2</u>
strongly agree	1	1
agree	17	15
neutral	10	7
disagree	52	55
strongly disagree	20	22
11. The pace of the computer was so slow that it made me impatient.	<u>session 1</u>	<u>session 2</u>
strongly disagree	6	4
disagree	34	37
neutral	18	12
agree	38	42
strongly agree	4	5
16. It was easy to read the words and questions on the screen.	<u>session 1</u>	<u>session 2</u>
strongly agree	1	2
agree	9	7
neutral	10	5
disagree	52	62
strongly disagree	28	24

Table 17

Response Percentages for the Test Satisfaction Factor  
for Items and Alternatives Over Both Sessions

<hr/>		
7. I felt that my performance on this test reflected my true knowledge of vocabulary.		
	<u>session 1</u>	<u>session 2</u>
strongly disagree	11	16
disagree	53	49
neutral	11	15
agree	19	19
strongly agree	6	1
17. I felt that the test did a good job of measuring my ability in vocabulary.		
	<u>session 1</u>	<u>session 2</u>
strongly agree	16	15
agree	49	47
neutral	19	24
disagree	14	12
strongly disagree	2	2
<hr/>		

Table 18

Response Percentages for the Time Pressure Factor  
for Items and Alternatives Over Both Sessions

<hr/>		
2. I felt less time pressure while taking this test than while taking conventional vocabulary tests.		
	<u>session 1</u>	<u>session 2</u>
strongly agree	1	1
agree	9	9
neutral	11	8
disagree	48	61
strongly disagree	31	21
4. The computer terminal made me feel that I had to answer the items as quickly as possible.		
	<u>session 1</u>	<u>session 2</u>
strongly agree	0	1
agree	4	5
neutral	6	3
disagree	53	64
strongly disagree	37	27
<hr/>		

Most of the correlation analyses between the attitude scale results and other variables such as ability levels and test length, in items and time, turned out to be negligible. However, there were a few exceptions. For instance, items 6 and 9 from the perceived test difficulty correlated moderately with the number of items correct (+.32) and Rasch ability (+.38), although item 9 was a particularly poor item. In addition, item 15 dealing with stress felt on the test correlated +.36 with number of items correct, indicating wrong answers accompanied higher stress levels and vice versa. Also, item 8 which concerned eye discomfort when viewing the CRT screen correlated +.34 with screen display color. This result showed a tendency for increased eye discomfort to be associated with the CRT screen with the white printing on black background display. Finally, items 12 and 14 from the perceived difficulty factor were moderately correlated (-.37 and -.43, respectively) with traditional exam scores for the course in which they were enrolled. These latter correlations indicated that students who perceived themselves to have performed poorly on the tailored vocabulary test tended to have done well on the course exams.

It is important to recognize that the correlations discussed above were just a few of a much larger number of correlation coefficients calculated. As such, caution must be exercised in regard to interpretations since the correlations could merely represent chance variations in the data. Therefore no great importance should be attached to the relationships discussed in this part of attitude results section.

The results of the MANOVA's to determine differences in attitude across test sessions and CRT screen display colors were both non-significant. The former result shows that there were no significant attitude changes toward the various aspects of the tailored testing situation from one session to the next. The latter result indicated that CRT screen display mode (black-on-white versus white-on-black) also did not significantly affect attitudes toward the tailored tests. These results held up regardless of whether the MANOVA's were run on raw scores or factor scores derived from the attitude scale responses.

#### Summary and Discussion

The overall purpose of the present research was to compare tailored testing procedures based on two prominent logistic models on the basis of the results of a live tailored testing study. Since both the one- and three-parameter models assume the measurement of a single latent trait dimension, it was decided to employ an item pool of vocabulary items with one primary factor for use in the comparison study. The results of the research could then be considered as basic groundwork for examining multidimensional tests, such as achievement tests, at a later date. In order to evaluate the two models, the following comparisons were undertaken: (a) the goodness of fit of the models, (b) the reliabilities of the two model-based tailored tests, (c) the ability estimates yielded by the two procedures, (d) the correlation of the ability estimates with the same outside criteria, (e) descriptive statistics for each procedure, (f) the convergence rates of the two methods, and (g) the information functions for the two types of tailored tests.

The most important finding of the study was the result that the three-parameter tailored tests were more informative than the one-parameter tailored tests. Moreover, the information function for the three-parameter tailored tests exceeded that for the 30 item traditional paper-and-pencil vocabulary test for the range of abilities in which most of the examinees were concentrated. However, in no case did the one-parameter tailored tests provide more information than the traditional test. The relative efficiency comparisons of the two types of tailored tests versus the traditional test shown in Figure 1 clearly illustrate the superiority of the three-parameter method in regard to tailored test information functions.

Another significant finding of the study was that the three-parameter tailored tests were more reliable than the one-parameter tailored tests. Unfortunately this result was not as straightforward as the test information result due to the presence of substantial nonconvergence of ability estimation for the three-parameter tailored testing procedures. Approximately one-third of the cases had to be removed from the reliability data analysis as a result of nonconvergence caused by the excessive difficulty of the item pool for these examinees.

However, the decision of whether or not to delete nonconvergence cases is debatable. On the one hand, advocates of the one-parameter model for tailored testing could argue that the method is robust with respect to item pools of inappropriate difficulty for the examinees, since nonconvergence never occurs. On the other hand, one could argue that nonconvergence does not occur with the three-parameter model if reasonable care is taken to insure the use of item pools of appropriate difficulty. In addition, it is quite likely that an empirical solution to the nonconvergence problem can be found so that abilities can be estimated even with item pools that are too difficult for examinees.

Samejima (1973) anticipated the nonconvergence problem with the three-parameter model and proposed a solution which has yet to be tried out by the present authors. However, work is currently in progress to attempt to develop an empirical "fix" of the nonconvergence problem, with several possibilities being explored. For example, one solution investigated is to arbitrarily estimate the examinee's ability at a fixed stepsize .693 less than the previous ability estimate in order to select the next item to be administered. Another method is to arbitrarily jump down to a very low ability (say -4.00) and hold it there until the responses to successive items permit an ability estimate to be calculated. A third approach is to use a variable stepsize downward in which the ability estimate first jumps down a full .693 stepsize, then if the item is still answered incorrectly, the ability estimate goes down half of a full stepsize, then half of that, etc. Early results of these techniques are encouraging and we are optimistic that the nonconvergence problem can be solved empirically.

A third finding of importance in the present study was the result that the three-parameter logistic model was superior to the one-parameter logistic model in terms of the goodness of fit criterion, namely mean squared deviations (MSD) of observed from predicted response data. Although the sampling distribution of the MSD statistic was unknown, previous research had shown the distribution to be approximately normal (Reckase, 1977). Thus the t-test results may be interpreted for this data as evidence that the three-parameter tailored testing procedure did a significantly better

job of fitting the response data than the one-parameter test. The result showed a closer match between the item responses predicted by the model and the actual observed responses for the three-parameter tailored tests than for the one-parameter tailored tests.

The results of the ability estimate intercorrelations, as well as the correlations of ability estimates with outside criteria, were inconclusive as far as comparing the two models were concerned. In general the consistent, moderately high degree of intercorrelation among the ability estimates yielded by both models over both sessions indicated that both procedures were measuring the same thing. Moreover, both of the tailored testing methods correlated equally well with the outside criteria measures. In this regard it should be noted that high correlations were not expected, since performance levels on a general vocabulary test would not necessarily lead to similar performances on course achievement tests. However, the achievement test scores were the only outside criteria available for the examinees.

In regard to the convergence rate comparison, the subjective analysis of the convergence plots on the whole indicated that the three-parameter tailored tests did a better job of arriving at stable ability estimates than the one-parameter tests. Of course, this result held only when 39 nonconvergence cases were removed from the data analysis. If included, the one-parameter tailored test convergence patterns would have been superior.

The descriptive statistics for the two tailored testing procedures showed the three-parameter tests to be slightly longer on the average, although test length differences would best be interpreted as being a function of the different item selection methods and stopping rules employed. Since the  $\pm .30$  acceptance range for the one-parameter method and the .70 information level cutoff for the three-parameter method were both somewhat arbitrary values derived from simulation and empirical studies, changes in these values would have changed the number of items administered. Both procedures functioned well on the average in administering items of appropriate difficulty for the examinees' estimated abilities during the tailored tests.

The attitude scale constructed to measure the students' attitudes toward tailored testing was found to be moderately reliable (with values ranging from .57 to .67), and the factor structures from the two testing sessions were found to be similar. However, the factor analysis results showed that several scale items shifted factors from one session to the next. In addition, item 9 and perhaps a few other items on the scale should be either deleted or revised due to their low or negative discrimination indexes. With these modifications, the reliability of the scale should improve considerably for future administrations.

The results from the attitude scale response data were generally favorable toward tailored testing. For instance, very few students were nervous or anxious about the prospect of taking a test on a CRT terminal instead of a paper-and-pencil format. However, most examinees felt considerable stress while taking the test itself, indicating some degree of involvement or motivation during testing. In terms of perceived test difficulty, most of the students found both that some items were too easy and that some were too difficult. In general they felt that they did as well on the tailored test

as other students, as well as on other vocabulary tests, and also that the tailored test did a good job of measuring their vocabulary abilities. In addition, their responses showed that the examinees were motivated to try to select the best answer to each item and that the pace of the test was comfortable.

On the negative side, most students felt that they could have performed better on the test if they had tried harder. They also indicated difficulty in reading the items on the CRT screen. Surprisingly, they felt the tailored test to have more time pressure than conventional tests, even though they had as much time as desired to respond to each item. Finally, the examinees did not feel that the tailored test results reflected their "true" vocabulary knowledge.

In some cases the attitude responses within a specific factor were directly contradictory. For example, in the perceived difficulty factor, responses for the first four items listed in Table 13 showed that the students found the vocabulary items to be easy and felt that they did well compared to other people and other vocabulary tests. Yet item 14 responses indicated exactly the opposite since most students did not feel confident of having done well on the test. And in fact, the ability estimate distributions pictured in Figures C-1, C-2, C-3, and C-4 in Appendix C reveal that the tailored vocabulary tests were difficult for the examinees compared to the calibration sample. The nonconvergence problem was another indication of the high test difficulty. It is difficult to reconcile such contradictions in the attitude scale response data, unless one argues that the examinees were randomly responding to the attitude items without much thought as to their choices.

The MANOVA results dealing with attitude changes over test sessions as well as CRT screen display mode were somewhat surprising in that both were nonsignificant. That is, attitude changes were expected in regard to anxiety in particular, since many examinees appeared to be nervous when encountering the CRT terminal for the first time. Perhaps the nonthreatening nature of the vocabulary test nullified any unusual anxiety levels, which might not have been the case if the test had counted toward course grades. Another expected result based on previous research (Koch & Patience, 1977) that did not materialize in the present study was different attitudes toward the two modes of CRT screen display. The results of the present study failed to corroborate the previous finding that the black-on-white mode was preferred to the white-on-black display. However, the correlation analyses results previously discussed did provide some evidence in this direction.

Previous attitude research (Koch & Patience, 1977) had failed to show any significant correlations between examinees' attitudes and performance on tailored tests. The present study again indicated that there was no significant linear relationship between ability estimates yielded by the tailored tests and attitude responses by the examinees. Even though such variables as motivation and anxiety might be expected to interact with test performance, no evidence to this effect could be found in the data from the present research.

Conclusion

A live tailored testing study was conducted to compare the results of using procedures based on either the one-parameter or the three-parameter logistic models to measure the performance of college students on multiple choice vocabulary items. Test items were selected for administration based on the information function, and maximum likelihood ability estimation was employed. The results of the study showed the three-parameter tailored testing procedure to be superior to the one-parameter procedure on the basis of test information, test-retest reliability, goodness of fit of observed to predicted item responses, and convergence rates to stable ability estimates.

No differences were found in the prediction of outside criteria. However, implicit in these results was the assumption that the nonconvergence problem encountered in one-third of the cases for the three-parameter procedure could be solved. Thus, based on the data reported in this study, the three-parameter tailored testing method was deemed the technique of choice, at least for unidimensional tests consisting of multiple choice items where guessing is a factor. The attitude results were generally favorable toward tailored testing, although no interaction was found between examinees' attitudes and test performance.

References

- Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.
- English, R. A., Reckase, M. D., & Patience, W. M. Application of tailored testing to achievement measurement. Behavior Research Methods & Instrumentation, 1977, 9, 158-161.
- Hambleton, R. K. An empirical investigation of the Rasch test theory model. Unpublished doctoral dissertation, University of Toronto, 1969.
- Hambleton, R. K., & Traub, R. E. Information curves and efficiency of three logistic test models. British Journal of Mathematical and Statistical Psychology, 1971, 24, 273-281.
- Ireland, C. M. An application of the Rasch one-parameter logistic model to individual intelligence testing in a tailored testing environment. Unpublished doctoral dissertation, University of Missouri - Columbia, 1976.
- Jensema, C. J. The validity of Bayesian tailored testing. Educational and Psychological Measurement, 1974, 34, 757-766.
- Koch, W. R., & Patience, W. M. Student attitudes toward tailored testing. Paper presented at the Second Conference on Computerized Adaptive Testing, Minneapolis, Minnesota, July, 1977.
- Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.
- Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper and Row, 1970.
- Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.
- Marco, G. L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14, 139-160.
- Monge, R. H. & Gardner, E. F. A program of research in adult differences in cognitive performance and learning: background for adult education and vocational retraining. (Final Report, Project No. 6-1963, Grant No. OEG 1-7-061963-0149). Syracuse, New York: Syracuse University, Department of Psychology, January 1972.
- Patience, W. M. Description of components in tailored testing. Behavior Research Methods & Instrumentation, 1977, 9, 153-157.
- Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

- Reckase, M. D. Development and application of a multivariate logistic latent trait model. (Doctoral dissertation, Syracuse University, 1972). Dissertation Abstracts International, 1973, 33. (University Microfilms No. 73-7762).
- Reckase, M. D. An interactive computer program for tailored testing based on the one-parameter logistic model. Behavior Research Methods & Instrumentation, 1974, 6, 208-212.
- Reckase, M. D. The effects of item pool characteristics on the operation of a tailored testing procedure. Paper presented at the spring meeting of the Psychometric Society, Murray Hill, New Jersey, 1976.
- Reckase, M. D. Ability estimation and item calibration using the one and three parameter logistic models: a comparative study. (Research Report 77-1). Columbia, Missouri. University of Missouri, Educational Psychology Department, November 1977. (AD A047943).
- Rentz, R. R., & Bashaw, W. L. The National Reference Scale for reading: an application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-179.
- Samejima, F. A comment on Birnbaum's three-parameter logistic model in the latent trait theory. Psychometrika, 1973, 38, 221-233.
- Urry, V. W. A Monte Carlo investigation of logistic mental test models. (Doctoral dissertation, Purdue University, 1970). Dissertation Abstracts International, 1971, 31, 6319B. (University Microfilms No. 71-9475).
- Urry, V. W. Tailored testing: a successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.
- Wood, R. L., Wingersky, M. S. & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. (ETS Research Memorandum RM-76-6). Princeton, New Jersey: Educational Testing Service, June 1976.
- Woodcock, R. W. Woodcock Reading Mastery Tests. Circle Pines, Minnesota: American Guidance Service, 1973.
- Wright, B. D. & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.
- Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

APPENDIX A

FIGURE A-1

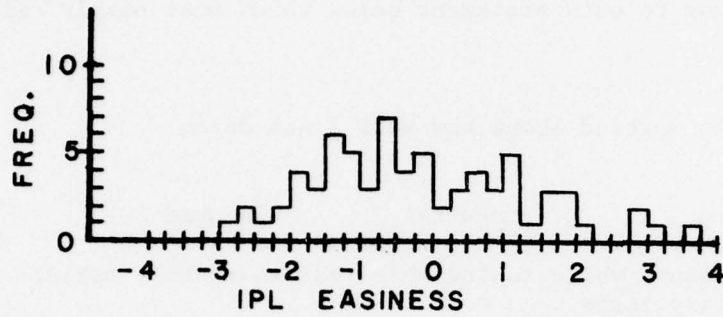


FIGURE A-2

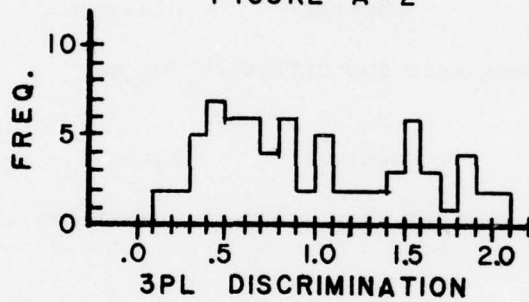


FIGURE A-3

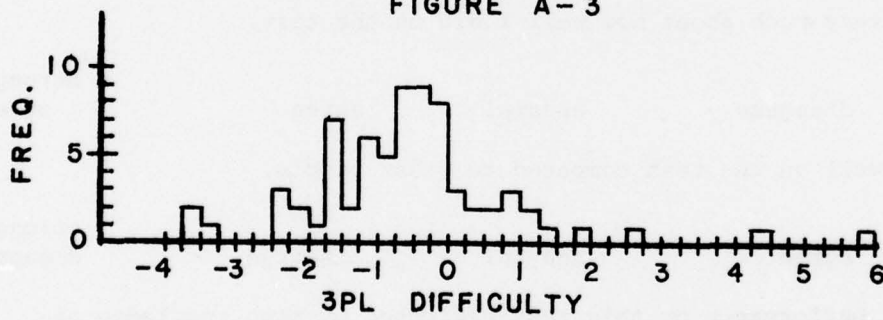
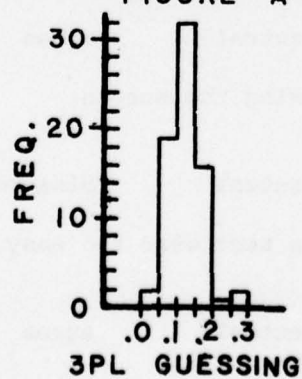


FIGURE A-4



APPENDIX B

ATTITUDE SURVEY

Please circle the response to each statement below which most nearly reflects your feelings or attitude.

1. During the test I was worried about how well I was doing.

strongly agree	agree	neutral	disagree	strongly disagree
-------------------	-------	---------	----------	----------------------

2. I felt less time pressure while taking this test than while taking conventional vocabulary tests.

strongly agree	agree	neutral	disagree	strongly disagree
-------------------	-------	---------	----------	----------------------

3. I felt that many of the items were too difficult for me.

strongly disagree	disagree	neutral	agree	strongly agree
----------------------	----------	---------	-------	-------------------

4. The computer terminal made me feel that I had to answer the items as quickly as possible.

strongly agree	agree	neutral	disagree	strongly disagree
-------------------	-------	---------	----------	----------------------

5. I didn't care very much about how well I did on the test.

strongly disagree	disagree	neutral	agree	strongly agree
----------------------	----------	---------	-------	-------------------

6. I think I did well on the test compared to other people.

strongly agree	agree	neutral	disagree	strongly disagree
-------------------	-------	---------	----------	----------------------

7. I felt that my performance on this test reflected my true knowledge of vocabulary.

strongly disagree	disagree	neutral	agree	strongly agree
----------------------	----------	---------	-------	-------------------

8. My eyes were uncomfortable when viewing the screen.

strongly agree	agree	neutral	disagree	strongly disagree
-------------------	-------	---------	----------	----------------------

9. I felt that many of the items on the test were too easy.

strongly disagree	disagree	neutral	agree	strongly agree
----------------------	----------	---------	-------	-------------------

10. I was nervous about coming here to take this test.

strongly agree	agree	neutral	disagree	strongly disagree
-------------------	-------	---------	----------	----------------------

11. The pace of the computer was so slow that it made me impatient.

strongly disagree	disagree	neutral	agree	strongly agree
----------------------	----------	---------	-------	-------------------

12. I feel that I did as well on this test as on other vocabulary tests I've taken.

strongly agree	agree	neutral	disagree	strongly disagree
-------------------	-------	---------	----------	----------------------

13. The computer terminal made me nervous.

strongly agree	agree	neutral	disagree	strongly disagree
-------------------	-------	---------	----------	----------------------

14. I felt confident that I did well on the test.

strongly disagree	disagree	neutral	agree	strongly agree
----------------------	----------	---------	-------	-------------------

15. I felt considerable stress while taking the test.

strongly disagree	disagree	neutral	agree	strongly agree
----------------------	----------	---------	-------	-------------------

16. It was easy to read the words and questions on the screen.

strongly agree	agree	neutral	disagree	strongly disagree
-------------------	-------	---------	----------	----------------------

17. I felt that the test did a good job of measuring my ability in vocabulary.

strongly agree	agree	neutral	disagree	strongly disagree
-------------------	-------	---------	----------	----------------------

18. I think I could have done better on the test if I had tried harder.

strongly disagree	disagree	neutral	agree	strongly agree
----------------------	----------	---------	-------	-------------------

19. I was careful to try to select the best answer to each question.

strongly disagree	disagree	neutral	agree	strongly agree
----------------------	----------	---------	-------	-------------------

20. I tried to finish the test quickly just to receive my 5 points credit.

strongly agree	agree	neutral	disagree	strongly disagree
-------------------	-------	---------	----------	----------------------

# APPENDIX C

FIGURE C-1

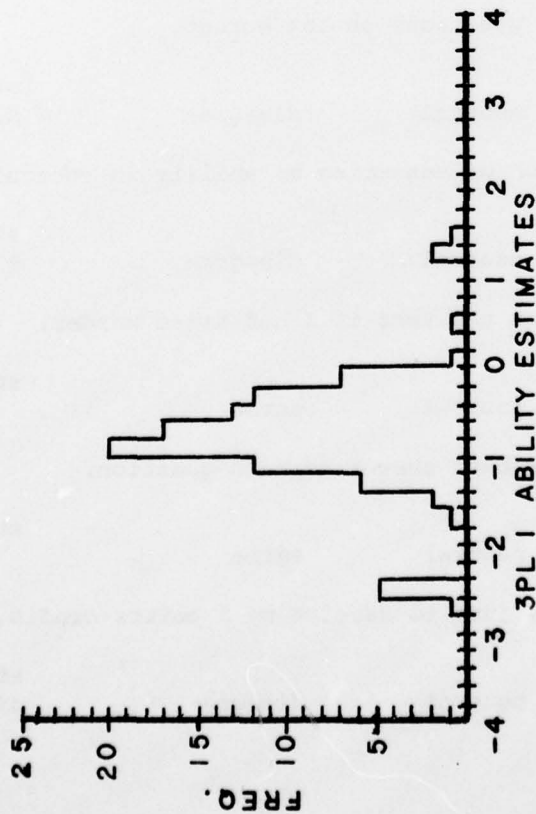


FIGURE C-2

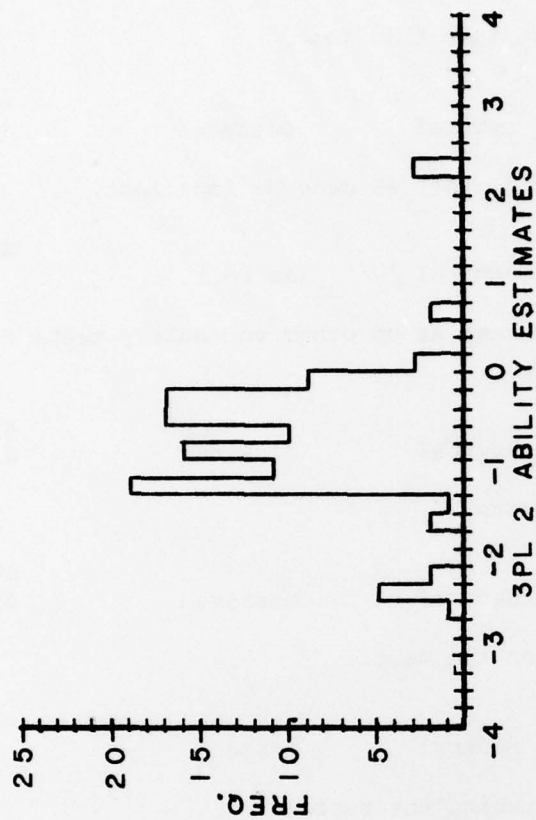


FIGURE C-3

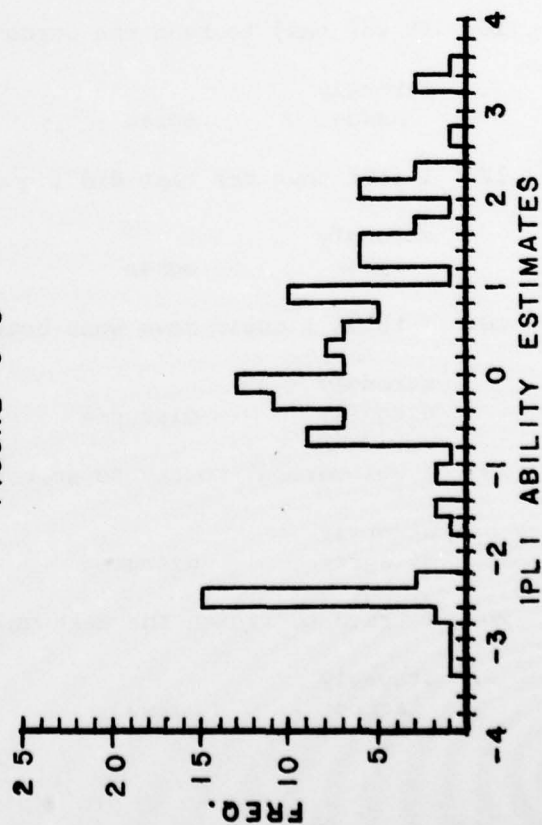
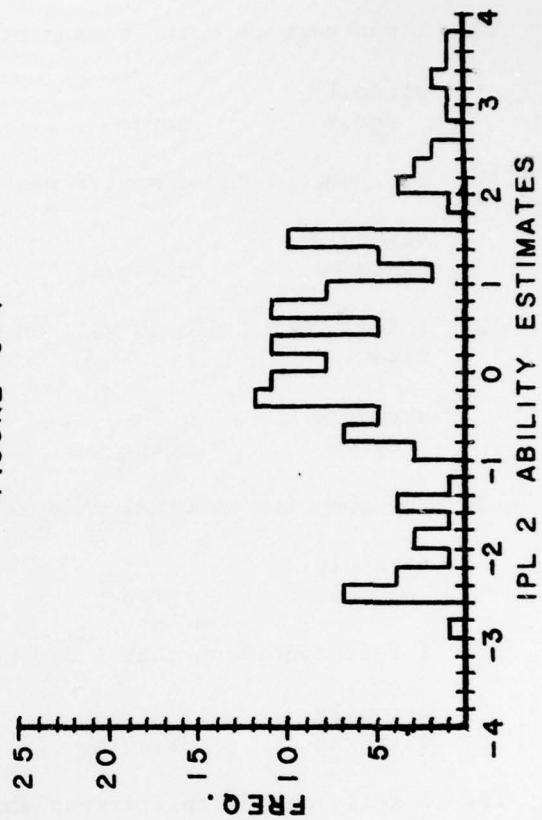


FIGURE C-4



Navy

- 1 DR. JACK ADAMS  
OFFICE OF NAVAL RESEARCH BRANCH  
223 OLD MAPLEHURST ROAD  
LONDON, NW, 15TH ENGLAND
- 1 Dr. Ed Aiken  
Navy Personnel R&D Center  
San Diego, CA 92152
- 1 Dr. Jack R. Forsting  
Provost & Academic Dean  
U.S. Naval Postgraduate School  
Monterey, CA 93940
- 1 DR. JOHN F. PROCK  
NAVY PERSONNEL R&D CENTER  
SAN DIEGO, CA 92152
- 1 DR. MAURICE CALLAHAN  
NODAC (CODE 2)  
DEPT. OF THE NAVY  
BLDG. 2, WASHINGTON NAVY YARD  
(ANACOSTIA)  
WASHINGTON, DC 20374
- 1 Dept. of the Navy  
CHNAVMAT (NMAT 034D)  
Washington, DC 20350
- 1 Chief of Naval Education and  
Training Support )-(01A)  
Pensacola, FL 32509
- 1 Dr. Charles E. Davis  
ONR Branch Office  
536 S. Clark Street  
Chicago, IL 60605
- 1 Mr. James S. Duva  
Chief, Human Factors Laboratory  
Naval Training Equipment Center  
(Code N-215)  
Orlando, Florida 32813
- 5 Dr. Marshall J. Farr, Director  
Personnel & Training Research Programs  
Office of Naval Research (Code 458)  
Arlington, VA 22217

Navy

- 1 DR. PAT FEDERICO  
NAVY PERSONNEL R&D CENTER  
SAN DIEGO, CA 92152
- 1 CDR John Ferguson, MSC, USN  
Naval Medical R&D Command (Code 44)  
National Naval Medical Center  
Bethesda, MD 20014
- 1 Dr. Eugene E. Gloye  
ONR Branch Office  
1030 East Green Street  
Pasadena, CA 91101
- 1 CAPT. D.M. GRAGG, MC, USN  
HEAD, SECTION ON MEDICAL EDUCATION  
UNIFORMED SERVICES UNIV. OF THE  
HEALTH SCIENCES  
6917 ARLINGTON ROAD  
BETHESDA, MD 20014
- 1 MR. GEORGE N. GRAINE  
NAVAL SEA SYSTEMS COMMAND  
SEA 047C112  
WASHINGTON, DC 20362
- 1 Dr. Steve Harris  
Code L522  
NAMRL  
Pensacola FL 32508
- 1 CDR Robert S. Kennedy  
Naval Aerospace Medical and  
Research Lab  
Box 29407  
New Orleans, LA 70139
- 1 Dr. Norman J. Kerr  
Chief of Naval Technical Training  
Naval Air Station Memphis (75)  
Millington, TN 38054
- 1 CHAIRMAN, LEADERSHIP & LAW DEPT.  
DIV. OF PROFESSIONAL DEVELOPMENT  
U.S. NAVAL ACADEMY  
ANNAPOLIS, MD 21402

## Navy

- 1 Dr. James Lester  
ONR Branch Office  
495 Summer Street  
Boston, MA 02210
- 1 Dr. James McFride  
Code 301  
Navy Personnel R&D Center  
San Diego, CA 92152
- 1 DR. WILLIAM MONTAGUE  
NAVY PERSONNEL R&D CENTER  
SAN DIEGO, CA 92152
- 1 Dr. Robert Morrison  
Code 301  
Navy Personnel R&D Center  
San Diego, CA 92152
- 1 Commanding Officer  
U.S. Naval Amphibious School  
Coronado, CA 92155
- 1 Commanding Officer  
Naval Health Research  
Center  
Attn: Library  
San Diego, CA 92152
- 1 CDR PAUL NELSON  
NAVAL MEDICAL R&D COMMAND  
CODE 44  
NATIONAL NAVAL MEDICAL CENTER  
BETHESDA, MD 20014
- 1 DR. RICHARD J. NIEHAUS  
CODE 301  
OFFICE OF CIVILIAN PERSONNEL  
NAVY DEPT  
WASHINGTON, DC 20390
- 1 Library  
Navy Personnel R&D Center  
San Diego, CA 92152
- 6 Commanding Officer  
Naval Research Laboratory  
Code 2627  
Washington, DC 20390

## Navy

- 1 OFFICE OF CIVILIAN PERSONNEL  
(CODE 26)  
DEPT. OF THE NAVY  
WASHINGTON, DC 20390
- 1 JOHN OLSEN  
CHIEF OF NAVAL EDUCATION &  
TRAINING SUPPORT  
PENSACOLA, FL 32509
- 1 Office of Naval Research  
Code 437  
800 N. Quincy Street  
Arlington, VA 22217
- 1 Office of Naval Research  
Code 441  
800 N. Quincy Street  
Arlington, VA 22217
- 1 Scientific Director  
Office of Naval Research  
Scientific Liaison Group/Tokyo  
American Embassy  
APO San Francisco, CA 96503
- 1 SCIENTIFIC ADVISOR TO THE CHIEF  
OF NAVAL PERSONNEL  
NAVAL BUREAU OF PERSONNEL (PERS OR)  
RM. 4410, ARLINGTON ANNEX  
WASHINGTON, DC 20370
- 1 DR. RICHARD A. POLLAK  
ACADEMIC COMPUTING CENTER  
U.S. NAVAL ACADEMY  
ANNAPOLIS, MD 21402
- 1 Mr. Arnold I. Rubinstein  
Human Resources Program Manager  
Naval Material Command (0344)  
Room 1044, Crystal Plaza #5  
Washington, DC 20360
- 1 A. A. SJOPOLM  
TECH. SUPPORT, CODE 201  
NAVY PERSONNEL R&D CENTER  
SAN DIEGO, CA 92152

## Navy

- 1 Mr. Robert Smith  
Office of Chief of Naval Operations  
OP-987E  
Washington, DC 20350
- 1 CDR Charles J. Theisen, JR. MSC, USN  
Head Human Factors Engineering Div.  
Naval Air Development Center  
Warminster, PA 18974
- 1 W. Gary Thomson  
Naval Ocean Systems Center  
Code 7132  
San Diego, CA 92152
- 1 DR. H.M. WEST III  
DEPUTY ADCNO FOR CIVILIAN PLANNING  
AND PROGRAMMING  
RM. 2625, ARLINGTON ANNEX  
WASHINGTON, DC 20370
- 1 DR. MARTIN F. WISKOFF  
NAVY PERSONNEL R & D CENTER  
SAN DIEGO, CA 92152

## Army

- 1 ARI Field Unit-Leavenworth  
P.O. Box 3122  
Ft. Leavenworth, KS 66027
- 1 HO USAAREUE & 7th Army  
ODCSOPS  
USAAREUE Director of GED  
APO New York 09403
- 1 Commandant  
U.S. Army Infantry School  
Ft. Benning, GA 31905  
Attn: ATSH-I-V-IT (Cpt. Pinton)
- 1 DR. RALPH CANTER  
U.S. ARMY RESEARCH INSTITUTE  
5001 EISENHOWER AVENUE  
ALEXANDRIA, VA 22333
- 1 DR. RALPH DUSEK  
U.S. ARMY RESEARCH INSTITUTE  
5001 EISENHOWER AVENUE  
ALEXANDRIA, VA 22333
- 1 Dr. A. Hyman  
Army Research Institute  
5001 Eisenhower Blvd.  
Alexandria, VA 22333
- 1 Dr. Ed Johnson  
Army Research Institute  
5001 Eisenhower Blvd.  
Alexandria, VA 22333
- 1 Dr. Milton S. Katz  
Individual Training & Skill  
Evaluation Technical Area  
U.S. Army Research Institute  
5001 Eisenhower Avenue  
Alexandria, VA 22333
- 1 DR. JAMES L. RANEY  
U.S. ARMY RESEARCH INSTITUTE  
5001 EISENHOWER AVENUE  
ALEXANDRIA, VA 22333

Army

- 1 Director, Training Development  
U.S. Army Administration Center  
ATTN: Dr. Sherrill  
Ft. Benjamin Harrison, IN 46218
- 1 Dr. Joseph Ward  
U.S. Army Research Institute  
5001 Eisenhower Avenue  
Alexandria, VA 22333

Air Force

- 1 Air Force Human Resources Lab  
AFHRL/PED  
Brooks AFB, TX 78235
- 1 Air University Library  
AUL/LSE 76/443  
Maxwell AFB, AL 36112
- 1 CDR. MERCER  
CNET LIAISON OFFICER  
AFHRL/FLYING TRAINING DIV.  
WILLIAMS AFB, AZ 85224
- 1 Dr. Donald E. Meyer  
U.S. Air Force  
ATC/XPTD  
Randolph AFB, TX 78148
- 1 Personnel Analysis Division  
HQ USAF/DPXXA  
Washington, DC 20330
- 1 Research Branch  
AFMPC/DPMYP  
Randolph AFB, TX 78148
- 1 Dr. Marty Rockway (AFHRL/TT)  
Lovry AFB  
Colorado 80230
- 1 Major Wayne S. Sellman  
Chief, Personnel Testing  
AFMPC/DPNYPT  
Randolph AFB, TX 78148

Marines

- 1 Director, Office of Manpower Utilization 1  
HQ, Marine Corps (MPU)  
BCB, Bldg. 2009  
Quantico, VA 22134

- 1 DR. A.L. SLAFKOSKY  
SCIENTIFIC ADVISOR (CODE RD-1)  
HQ, U.S. MARINE CORPS  
WASHINGTON, DC 20380

CoastGuard

MR. JOSEPH J. COWAN, CHIEF  
PSYCHOLOGICAL RESEARCH (G-P-1/62)  
U.S. COAST GUARD HQ  
WASHINGTON, DC 20590

# Other DoD

- 1 Dr. Stephen Andriole  
ADVANCED RESEARCH PROJECTS AGENCY  
1400 WILSON BLVD.  
ARLINGTON, VA 22209
- 12 Defense Documentation Center  
Cameron Station, Bldg. 5  
Alexandria, VA 22314  
Attn: TC
- 1 Dr. Dexter Fletcher  
ADVANCED RESEARCH PROJECTS AGENCY  
1400 WILSON PLVD.  
ARLINGTON, VA 22209
- 1 Military Assistant for Human Resources  
Office of the Director of Defense  
Research & Engineering  
Room 3D129, the Pentagon  
Washington, DC 20301
- 1 Director, Research & Data  
OSD/MRA&L (Rm. 3E919)  
The Pentagon  
Washington, DC 20301
- 1 Mr. Fredrick W. Suffa  
MPP (A&R)  
2B269  
Pentagon  
Washington, D.C. 20301

# Civil Govt.

- 1 Dr. Susan Chipman  
Basic Skills Program  
National Institute of Education  
1200 19th Street NW  
Washington, DC 20208
- 1 Dr. Lorraine D. Eyde  
Personnel R&D Center  
U.S. Civil Service Commission  
1900 E Street NW  
Washington, D.C. 20415
- 1 Dr. William Gorham, Director  
Personnel R&D Center  
U.S. Civil Service Commission  
1900 E Street NW  
Washington, DC 20415
- 1 Dr. Joseph Markowitz  
Office of Research and Development  
Central Intelligence Agency  
Washington, DC 20205
- 1 Dr. Andrew R. Molnar  
Science Education Dev.  
and Research  
National Science Foundation  
Washington, DC 20550
- 1 Dr. H. Wallace Sinaiko, Director  
Manpower Research & Advisory Service  
Smithsonian Institution  
801 K. Pitt Street  
Alexandria, VA 22314
- 1 Robert W. Stump  
Education & Work Group  
National Institute of Education  
1200 19th Street NW  
Washington, DC 20208
- 1 Dr. Vern W. Hury  
Personnel R&D Center  
U.S. Civil Service Commission  
1900 E Street NW  
Washington, DC 20415

Civil Govt

- 1 C.S. WINIEWICZ  
U.S. CIVIL SERVICE COMMISSION  
REGIONAL PSYCHOLOGIST  
230 S. DEARBORN STREET  
CHICAGO, IL 60604
- 1 Dr. Joseph L. Young, Director  
Memory & Cognitive Processes  
National Science Foundation  
Washington, DC 20550

Non Govt

- 1 PROF. EARL A. ALLUISI  
DEPT. OF PSYCHOLOGY  
CODE 227  
OLD DOMINION UNIVERSITY  
NORFOLK, VA 23508
- 1 1 psychological research unit  
Dept. of Defense (Army Office)  
Campbell Park Offices  
Canberra ACT 2600, Australia
- 1 Ms. Carole A. Bagley  
Minnesota Educational Computing  
Consortium  
2520 Broadway Drive  
St. Paul, MN 55112
- 1 Dr. Gerald V. Parrett  
Dept. of Psychology  
University of Akron  
Akron, OH 44325
- 1 Dr. Nicholas A. Pond  
Dept. of Psychology  
Sacramento State College  
600 Jay Street  
Sacramento, CA 95819
- 1 Dr. David G. Fowers  
Institute for Social Research  
University of Michigan  
Ann Arbor, MI 48106
- 1 Dr. Robert K. Franson  
1A Tully Building  
Florida State Univ.  
Tallahassee, FL 32306
- 1 DR. C. VICTOR FUMDERSON  
WICAT INC.  
UNIVERSITY PLAZA, SUITE 10  
1160 SO. STATE ST.  
OREM, UT 84057
- 1 Dr. John P. Carroll  
Psychometric Lab  
Univ. of No. Carolina  
Davie Hall 012A  
Chapel Hill, NC 27514

Non Govt

- 1 Dr. A. Charnes  
EEB 203E  
University of Texas  
Austin, TX 78712
- 1 Dr. Kenneth E. Clark  
College of Arts & Sciences  
University of Rochester  
River Campus Station  
Rochester, NY 14627
- 1 Dr. Norman Cliff  
Dept. of Psychology  
Univ. of So. California  
University Park  
Los Angeles, CA 90007
- 1 Dr. Allen M. Collins  
Eolt Peranek & Newman, Inc.  
50 Moulton Street  
Cambridge, Ma 02138
- 1 Dr. John J. Collins  
Essex Corporation  
201 N. Fairfax Street  
Alexandria, VA 22314
- 1 Dr. Meredith Crawford  
5605 Montgomery Street  
Chevy Chase, MD 20015
- 1 DR. RENE V. DAVIS  
DEPT. OF PSYCHOLOGY  
UNIV. OF MINNESOTA  
75 E. RIVER RD.  
MINNEAPOLIS, MN 55455
- 1 Dr. Ruth Day  
Center for Advanced Study  
in Behavioral Sciences  
202 Junipero Serra Blvd.  
Stanford, CA 94305
- 1 Dr. Marvin D. Dunnette  
N492 Elliott Hall  
Dept. of Psychology  
Univ. of Minnesota  
Minneapolis, MN 55455

Non Govt

- 1 Dr. Richard L. Ferguson  
The American College Testing Program  
P.O. Box 168  
Iowa City, IA 52240
- 1 Dr. Victor Fields  
Dept. of Psychology  
Montgomery College  
Rockville, MD 20850
- 1 Dr. Edwin A. Fleishman  
Advanced Research Resources Organ.  
8555 Sixteenth Street  
Silver Spring, MD 20910
- 1 Dr. John P. Frederiksen  
Polt Peranek & Newman  
50 Moulton Street  
Cambridge, MA 02138
- 1 Dr. Vernon S. Gerlach  
College of Education  
146 Payne Bldg. B  
Arizona State University  
Tempe, AZ 85281
- 1 DR. ROBERT GLASER  
LRDC  
UNIVERSITY OF PITTSBURGH  
3929 O'HARA STREET  
PITTSBURGH, PA 15213
- 1 DR. JAMES G. GREENO  
LRDC  
UNIVERSITY OF PITTSBURGH  
3929 O'HARA STREET  
PITTSBURGH, PA 15213
- 1 Dr. Ron Hambleton  
School of Education  
University of Massachusetts  
Amherst, MA 01002
- 2 Dr. Barbara Hayes-Roth  
The Rand Corporation  
1700 Main Street  
Santa Monica, CA 90406

NON GOVT

- 1 Dr. James R. Hoffman  
Department of Psychology  
University of Delaware  
Newark, DE 19711
- 1 HUMPRO/Ft. Knox office  
P.O. Box 293  
Ft. Knox, KY 40121
- 1 Mr. Gary Irving  
Data Sciences Division  
Technology Services Corporation  
2811 Wilshire Blvd.  
Santa Monica CA 90403
- 1 DR. LAWRENCE B. JOHNSON  
LAWRENCE JOHNSON & ASSOC., INC.  
SUITE 502  
2001 S STREET NW  
WASHINGTON, DC 20009
- 1 Dr. Roger A. Kaufman  
203 Dodd Hall  
Florida State Univ.  
Tallahassee, FL 32306
- 1 Dr. Steven W. Keele  
Dept. of Psychology  
University of Oregon  
Eugene, OR 97403
- 1 Mr. Marlin Kroger  
1117 Via Goleta  
Palos Verdes Estates, CA 90274
- 1 LCOL. C.R.J. LAFLEUR  
PERSONNEL APPLIED RESEARCH  
NATIONAL DEFENSE HQS  
101 COLONEL BY DRIVE  
OTTAWA, CANADA K1A 0K2
- 1 MR. W.E. LASSITER  
DATA SOLUTIONS CORP.  
2095 CHAIN BRIDGE ROAD  
VIENNA, VA 22180
- 1 Dr. Frederick M. Lord  
Educational Testing Service  
Princeton, NJ 08540

Non Govt

- 1 Dr. Robert R. Mackie  
Human Factors Research, Inc.  
6780 Cortona Drive  
Santa Barbara Research Pk.  
Goleta, CA 93017
- 1 Dr. William C. Mann  
USC-Information Sciences Inst.  
4676 Admiralty Way  
Marina del Rey, CA 90291
- 1 Mr. Edmond Marks  
304 Grange Bldg.  
Pennsylvania State Univ.  
University Park, PA 16802
- 1 Dr. Donald A Norman  
Dept. of Psychology C-009  
Univ. of California, San Diego  
La Jolla, CA 92093
- 1 Dr. Melvin R. Novick  
Iowa Testing Programs  
University of Iowa  
Iowa City, IA 52242
- 1 Dr. Jesse Orlansky  
Institute for Defense Analysis  
400 Army Navy Drive  
Arlington, VA 22202
- 1 Mr. A. J. Pesch, President  
Eclotech Associates, Inc.  
P. O. Box 178  
N. Stonington, CT 06359
- 1 DR. STEVEN M. PJNE  
4950 Douglas Avenue  
Golden Valley, MN 55416
- 1 DR. DJANE M. RAMSEY-KLEE  
R-K RESEARCH & SYSTEM DESIGN  
3947 RIDGEMONT DRIVE  
MALIBU, CA 90265

Non Govt

- 1 MIN. RET. M. RAUCH  
P II 4  
BUNDESMINISTERIUM DER VERTEIDIGUNG  
POSTFACH 161  
53 BONN 1, GERMANY
- 1 Dr. Mark D. Reckase  
Educational Psychology Dept.  
University of Missouri-Columbia  
12 Hill Hall  
Columbia, MO 65201
- 1 Dr. Joseph W. Rigney  
Univ. of So. California  
Behavioral Technology Labs  
3717 South Hope Street  
Los Angeles, CA 90007
- 1 Dr. Leonard L. Rosenbaum, Chairman  
Department of Psychology  
Montgomery College  
Rockville, MD 20850
- 1 PROF. FUMIKO SAMEJIMA  
DEPT. OF PSYCHOLOGY  
UNIVERSITY OF TENNESSEE  
KNOXVILLE, TN 37916
- 1 Dr. Benjamin Schneider  
Dept. of Psychology  
Univ. of Maryland  
College Park, MD 20742
- 1 Dr. Lyle Schoenfeldt  
School of Management  
Rensselaer Polytechnic Institute  
Troy, NY 12181
- 1 Dr. Robert Singer, Director  
Motor Learning Research Lab  
Florida State University  
212 Montgomery Gym  
Tallahassee, FL 32306
- 1 Dr. Robert Sternberg  
Dept. of Psychology  
Yale University  
Box 11A, Yale Station  
New Haven, CT 06520

Non Govt

- 1 Cr. C. Harold Stone  
1428 Virginia Avenue  
Glendale, CA 91202
- 1 Mr. D. J. Sullivan  
c/o Canyon Research Group, Inc.  
741 Lakefield Road  
Westlake Village, CA 91361
- 1 DR. PATRICK SUPPES  
INSTITUTE FOR MATHEMATICAL STUDIES  
THE SOCIAL SCIENCES  
STANFORD UNIVERSITY  
STANFORD, CA 94305
- 1 Dr. Kikumi Tatsuoka  
Computer Based Education Research  
Laboratory  
252 Engineering Research Laboratory  
University of Illinois  
Urbana, IL 61801
- 1 Dr. Fenton J. Underwood  
Dept. of Psychology  
Northwestern University  
Evanston, IL 60201
- 1 Dr. Willard S. Vaughan, Jr.  
Oceanautics, Inc.  
422 Sixth Street  
Annapolis, MD 21403
- 1 Dr. Robert Vineberg  
HUMRRO/Western Division  
27857 Perwick Drive  
Carmel, CA 93921
- 1 Dr. John Wennous  
Department of Management  
Michigan University  
East Lansing, MI 48824
- 1 Dr. David J. Weiss  
N660 Elliott Hall  
University of Minnesota  
75 E. River Road  
Minneapolis, MN 55455

Non Govt

1 Dr. Anita West  
Denver Research Institute  
University of Denver  
Denver, CO 80201

1 DR. SUSAN E. WHITELY  
PSYCHOLOGY DEPARTMENT  
UNIVERSITY OF KANSAS  
LAWRENCE, KANSAS 66044